ISSN 2395-8618

ISSN 2395-8618

ISSN 2395-8618

Hybrid Entity Driven News Detection on Twitter

Linn Vikre, Henning M. Wold, Özlem Özgöbek, and Jon Atle Gulla

***Abstract*—In recent years, Twitter has become one of the most popular microblogging services on the Internet. People sharing their thoughts and feelings as well as the events happening around them, makes Twitter a promising source of the most recent news received directly from the observers. But detecting the newsworthy tweets is a challenging task. In this paper we propose a new hybrid method for detecting real-time news on Twitter using locality-sensitive hashing (LSH) and named-entity recognition (NER). The method is tested on 72,000 tweets from the San Fransisco area and yields a precision of 0.917.**

I. Introduction

Twitter[1] is a popular social media platform where users can share their opinions and publish facts about everything from news to more personal matters. For example in an emergency situation, people can provide information about the situation as observers, or give relevant knowledge about the situation obtained by other sources and then share it through Twitter [1]. On the one hand, this gives Twitter great potential in being able to discover breaking news and follow all angles surrounding a novelty. On the other hand, it is easy to spread a rumor or unreliable facts using Twitter.

Focusing on news, there exists a plethora of different ways to convey the news depending on your location, your political standing, and so on. Therefore there are several factors that may influence how a tweet is formulated, and how it is interpreted. As Twitter is a fast growing social media platform with 302 million monthly active users and around 500 million tweets sent per day[2], it seems possible to discover information which is not yet published news around the world. Previous research [2], [3], [4] shows that it is possible to discover news on Twitter. It has, however, proven to be difficult to detect these news at early stages, namely before they are put on the newswire. The difficulty lies in the shortness of text allowed in tweets. A user on Twitter can only express themselves using 140 characters, which provokes the use of abbreviations, incorrect sentence structure, and language.

This paper explores how data pre-processing, named-entity recognition and locality-sensitive hashing (LSH) can be used to achieve better results when detecting breaking news. Our system is based on clustering tweets with similar content on the work of [5]. Their approach enables us to detect the topics that are being discussed on Twitter in a given period of time without using any Twitter-specific services, such as trending Topics[3]. Our base assumption is that if an event occurs that can be considered as a breaking news, that will lead to a "burst of activity" of people tweeting about it as described for general document streams in [6]. This will lead to a new, high activity, cluster appearing that contains tweets about the event. Keeping this in mind, we explore how the various parameters in the LSH implementation can be changed to increase the efficacy of our approach.

Central to this work is the ability to identify breaking news from a live feed of real-time tweets. The main motivation for this is that finding breaking news is of little value if they are detected too long after the news event in question happened, as this makes the likelihood of it already having been picked up by the newswire all but certain.

In this paper, we address the following issues:

- Evaluate to what extent LSH clustering can be applied for live news detection on Twitter.
- Cluster tweets based on their contents using our chosen locality-sensitive hashing (LSH) algorithm implementation, and see how it performs on random tweets.
- Evaluate if named-entity recognition (NER) can be used as a filtering method to detect news.

The structure of this paper is as follows. In Section II we present the previous research in the field. Section III describes the different methods used to conduct the evaluations. In Section IV we explain the details of our experiments like the data sets and tools we used. In Section V we explain the results of our experiments and finally in Section VI we discuss our results and give the conclusions.

II. Related work

In [7] it is stated that using traditional NER systems on tweets were insufficient for news detection. Thus in this work it is developed and trained a classifier based on annotated tweets to recognize named entities. The classifier trained handles *Location*, *Organization* and *Person* classes. Later in [8] it is performed a research using the same three classes, in addition to a *Product* class. In this approach it is used k-nearest neighbors as a supervised method.

Chiticariu et. al. [9] presents the language NERL (NER rule language) which is a high-level language to customize,

Manuscript received on August 21, 2016, accepted for publication on December 18, 2016, published on June 20, 2017.

The authors are with the Department of Computer and Information Science, NTNU, Trondheim, Norway (e-mail: {vikre, henninwo}@stud.ntnu.no, {ozlemo, jag}@idi.ntnu.no).

[1]https://twitter.com/

[2]https://about.twitter.com/company

[3]https://twitter.com/trendingtopics

ISSN 2395-8618

understand, and build rule-based named entity annotators across different domains. This work does not particularly focuses on Twitter but it is important to understand the difficulties of named-entity recognition.

Li et. al. [10] presents a novel 2-step unsupervised method for automatic discovery of emerging named entities, which could potentially be linked to news events such as crises. Their system, called TwiNER, leverages global context that are obtained from Wikipedia [4] and the Web N-Gram corpus, and ranks segments that are potentially true named entities. The system currently does not categorize the detected named entities. However, the approach only works with targeted Twitter streams, which means it is not usable with our approach as the tweets come from a variety of users.

As Ritter et. al. [11] addresses the issue of the noisy nature and incorrectness of language in tweets by re-building the NLP-pipeline consists of part-of-speech tagging (POS-tagging), chunking, and finally named-entity recognition. Their study shows an increase in accuracy and obtained a large (41%) reduction in error compared with the Stanford POS tagger. In their work it is also shown that the features generated by POS tagging and chunking gives a benefit to the segmenting of named entities. In our work, we utilize the system they created, though only the part that classifies named entities. We use this only as an additional filtering step in our system to detect breaking news.

Petrovic et. al. [5] presents a locality sensitive hashing (LSH) algorithm for "First story detection" on Twitter. According to this work, LSH is able to overcome some of the limitations of the traditional approaches, e.g. centroids and vectors in term space weighted with idf (inverse document frequency). Their method drastically reduces the number of comparisons a tweet needs to have in order to find the N nearest neighbors. This is crucial in a system that should work as a real-time service. Their work also shows that their system can find news events after analyzing an entire corpus. We utilize their approach in our system also, but we focus more on the activity in tweet clusters in smaller time periods, to detect if something out of the ordinary is happening. By doing this, we are able to detect news events in real-time.

Agarwal [12] presents an approach that takes a continuous stream of Twitter data, processes them to filter out tweets characterized as "noise" and get the informative ones. After this, they use the filtered tweets to detect and predict trending topics at an early stage. We use some of their findings when pre-process to disqualify tweets unlikely to be about a news event.

III. Methodology

In this section, we introduce the different methods we used to conduct the experiments where the details are explained in the next section.

[4]http://wikipedia.com

A. Locality sensitive hashing

Locality sensitive hashing (LSH) [13] is a technique for finding the nearest neighbor document in vector space utilizing vectors of random values and representing hyperplanes to generate hashes. This approach reduces the time and space complexity when finding the nearest neighbor.

LSH hashes the input items using k hyperplanes. Increasing the value of k decreases the probability of collision between non-similar items, while at the same time decreasing the probability of collision between nearest neighbors. To alleviate this, the implementation contains L different hash tables, each with independently chosen random hyperplanes. These hash tables are also called "buckets". This increases the probability of the item colliding with its nearest neighbor in at least one of the L hash tables. In other words there is a high probability that the item's nearest neighbor is contained in one of the buckets the item gets assigned to. LSH thus seek to achieve a maximum probability for collision on similar items. To be able to perform LSH on tweets, each tweet is converted to vector space, using tf-idf combined with Euclidean normalization.

There are several parameters that can be adjusted to change the results of LSH. In this paper we seek to find optimal values for these parameters with respect to detecting news, and filtering out as many of the non-relevant tweets as possible.

Our approach for LSH is based on the one described by [5]. We use the cosine similarity between document vectors $\vec{u}$ and $\vec{v}$ to decide the nearest neighbor:

$$cos(\theta) = \frac{\vec{u} \cdot \vec{v}}{||\vec{u}|| \cdot ||\vec{v}||}$$

Specifics of the implementation is elaborated on in Section IV-B.

B. Entropy

After a the tweets are clustered, it is possible to measure its Shannon entropy [14], which also is called information entropy. This is done by concatenating the text of all the tweets in a cluster into a single document. By doing this, the Shannon entropy is given by

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i)$$

where $P(X_i) = \frac{n_i}{N}$ is a tokens's probability of occurring in the document (found by dividing the number of times the token, n_i, appears in the document with how many tokens are in the document, N, in total).

Shannon entropy is a measure of how much information is contained within a document. Petrovic et. al. [5] suggests that by ignoring clusters with low entropy, it is possible to filter out many clusters that are filled with spam as they tend to have very low entropy.

Tweets are limited to 140 characters, and as Shannon entropy effectively use the length of the text as a factor (in N), longer tweets will be given a higher entropy than shorter

ISSN 2395-8618

tweets by default. For example, if a cluster contains five completely equal tweets of 50 characters each, and another cluster contains five completely equal tweets of 100 characters each, the latter will always have a higher entropy. For this reason, it would be unwise to set the entropy threshold too high (we could miss interesting stories that are simply too short), neither would setting it too low be wise (many stories containing spam and chatter would not be filtered out). The parameter governing the lowest entropy allowed in a cluster is *MIN_ENTROPY*.

C. Named-entity recognition (NER)

To detect breaking news, in addition to LSH, we examined if named-entity recognition can improve the results that we get by using LSH to filter tweets.

Named-entity recognition (NER) is a type of natural language processing (NLP). NLP refers to research that explore how computers can be utilized to understand, interpret, and manipulate natural languages, such as text and speech [15]. NER is a powerful tool for analyzing the deeper meaning behind sentences or longer sequences of words [8]. Thus NER is commonly understood as the task of identifying so-called named-entities in a text, such as organizations, products, locations, and persons.

A few different methods exist for performing NER on text documents: The rule-based method, the use of machine learning, and a hybrid between the two. The rule-based method has two extensive drawbacks; it lacks both portability and robustness. For this reason, machine learning has emerged as the better choice for coping with the problem [16], and is the method we have decided to utilize for our research.

D. Online processing of data

Since we are interested in detecting breaking news, we have to process the real-time data in a quick way. So we have to set up our system in such a way that it is quick enough to get fed data from the Twitter streaming API. We examine the clusters in set windows of time that has a length of *WINDOW_TIME_IN_SECONDS* seconds. We take the difference between the timestamp of the tweet currently being processed and the last tweet processed before the previous output. If this difference is greater than the duration of our window of time, we output information about the clusters matching our set parameters. As keeping every tweet previously processed in memory would quickly exhaust available resources, we only keep the message of the original tweet for a cluster in memory, in addition to the tweets added to it during the current window. A downside of this is that it also limits the amount of information available to decide whether a given cluster is news relevant or not.

IV. Experiment

In this section, we present the details of the experiment we have devised.

A. Data sets

Our data set consist of 72,000 tweets collected from the San Fransisco area using the Twitter streaming API[5]. The data was collected over a period of 30 hours from May 11 to May 12, 2015. The API provides extensive metadata for the tweets, most of which are not of interest to us in this experiment. We thus, to conserve storage space, strip most of it away, only keeping the tweet text itself, the tweet id, the user id of the poster, and the timestamp the tweet was posted. As previously mentioned, our system needs to work for tweets arriving in real-time. Despite this, we still chose to have a static data set for the experiments so that the same data would be used in each test.

B. Tools

Before hashing, the tweets were run through the NER engine. This put an extra field into the underlying JSON structure of the tweets where entities were detected called "entities". This field contained a list of all the entities detected in the tweets.

To perform the experiment, we require an implementation of the LSH algorithm. We have based our implementation on the ones outlined in [5] and [17]. Our implementation consists of two modules. The first module is the main LSH module and takes in a stream of tweets and outputs a stream of tweets augmented with information about their nearest neighbor (the id of the nearest neighbor, and its cosine similarity to that neighbor). The second module takes in a stream of tweets augmented with information about their nearest neighbor and outputs clusters of related tweets based on parameters discussed below.

We use the following static parameters in our LSH implementation:

- 13 bit hash values (k)
- 36 hash tables (L)
- 20 collisions per hash value
- 2000 previous tweets comparison

These values remain unchanged between experiments. In addition to these, we have parameters we adjust between experiments to find ideal values. These parameters, along with their initial values, can be found in table I.

The first step is tokenizing the tweets by splitting the text into tokens while removing punctuation. Additionally, tokens identified URLs, mentions or hashtags are not included in the list of tokens. If the tweet contains any non-ASCII tokens, the entire tweet is discarded. The entire tweet is also discarded if any of its tokens are included in the *IGNORE* list, or if the number of tokens left after tokenizing is $<$ *MIN_TOKENS*. After tokenizing, the nearest neighbor and the cosine similarity to this neighbor is found. Once this is done, the results are output and used in the next module responsible for creating clusters of related stories.

[5]http://dev.twitter.com

ISSN 2395-8618

The next module transforms the stream output from the LSH module into clusters of tweets based on a few parameters. If the cosine similarity between a tweet and its nearest neighbor is at least *MIN_COSSIM_FOR_EXISTING*, it is added to the same cluster as its nearest neighbor. If not, a new cluster is created with the tweet as its first message. After *WINDOW_SIZE_IN_SECONDS* seconds have passed between the timestamp of the first tweet processed and the current tweet being processed, all stories matching certain parameters will be printed. A cluster must have had at least *MIN_TWEETS_IN_SLICE* tweets added to it during the current window to be printed. Additionally at least *MIN_UNIQUES* unique users must have had their tweets added to that cluster for it to be printed. The last parameter to be matched for a cluster to be printed is that its information entropy must be at least *MIN_ENTROPY*.

In addition to finding optimal LSH parameters for detecting breaking news on Twitter, we want to examine if utilizing named-entity recognition can further improve the results. To this end we have used the work of [11]. They achieved better results than the baseline presented in their work to customize the named-entity recognition engine to fit the nature of tweets. We have thus elected to use the implementation the authors created, which they have published open source on the Internet[6].

C. Conducting the experiment

When conducting the experiment, we first need to establish a baseline. For this baseline, we gave the parameters the values listed in table I and calculated the precision achieved using those values. We define precision here as the proportion of the returned clusters deemed as news. The values of *MIN_COSSIM_FOR_ EXISTING* and *MIN_ENTROPY* in the baseline were chosen based on the findings of [5], while the values of *MIN_TOKENS* and *IGNORE* were chosen based on the findings of [12].

The *MIN_TWEETS_IN_SLICE* parameter is initially set to 1 (that is any non-empty cluster may get through the clustering algorithm). The reasoning behind this is that we suspect the value to have an impact on the results that we want to test, but setting it to 1 for the baseline effectively switches it off as to not pollute the results.

Another parameter of interest to the baseline is *MIN_UNIQUES*, which dictates how many unique users must have tweets in a cluster for it be to qualified. If this value is set lower than 2, it opens the proverbial floodgates for various single-user spam accounts, irrelevant or out-of-context tweets and other similar phenomena. We thus elected to set this value to 2 in the baseline and only examine values higher than this. As a news event is highly unlikely to be of interest if only one person is writing about it, with no one retweeting them, we consider this a fair choice to make.

[6]https://github.com/aritter/twitter_nlp

The final tested parameter is *WINDOW_SIZE_IN_SECONDS*, which dictates the length of time between clusters being printed. We were quite unsure of what would be the ideal values for this parameter. A short windows would enable the system to more rapidly detect news events. On the other hand, it is possible that more time has to elapse for the news events to become properly identifiable in the clusters. We thus defined the values to test to be from 5 minutes (300 seconds) to 25 minutes (1500 seconds). We set the baseline to be the middle of the range; 15 minutes (900 seconds).

Though the precision values calculated are valuable, another interesting facet is how the number of relevant clusters and non-relevant clusters change between the tests. We want to maximize the number of relevant clusters while minimizing the number of not relevant tweets.

After recording the number of relevant and not relevant clusters in the baseline, we proceeded to change the parameters one at a time, leaving the others at their baseline value. This enabled us to see how changing a parameter changes results. An overview over the various tests we performed, as well as their results, can be found in table II.

Next we used the information from the NER engine to further filter the clusters. This was done by simply dropping those not contain the "entities" field from the output. We did not perform this additional step on all tests; the ones we did perform it for are marked with NER in table II.

Finally, we combined a few of the most promising candidates for news detection in a test. This test is marked FINAL in table II.

V. Results

Our results, which can fully be found in Table II, show that the parameters we picked out for our baseline had a fairly average precision of 0.472. By combining various values for the different parameters, in addition to including named-entity recognition, we were able to increase this precision to 0.917.

From the results we can see that there were three things in particular that largely affected the result: the minimum amount of entropy allowed in a cluster (*MIN_ENTROPY* - ME), whether or not named entities were used to filter the cluster (NER), and the minimum cosine similarity between two tweets for them to be put in the same bucket (*MIN_COSSIM_FOR_EXISTING* - MCFE). We had assumed in advance that using NER to filter the clusters would have a positive impact. That this turns out to be the case is thus not surprising. The MCFE parameter is still within the range suggested for it in [5] for the optimal values we found. The experiments show that when MCFE is set lower than its suggested values, the precision value plummets. We did not test having this value set higher than their highest suggested value. The reasoning being that by increasing it, you increase the number of clusters and reduce the number of tweets landing in the same clusters. Meanwhile is the ME parameter set higher than the suggested value of the paper. This seems

ISSN 2395-8618

TABLE I
BASELINE PARAMETERS FOR FILTERING TWEETS AND THE LSH CLUSTERING

Baseline	
Parameter	**Value**
MIN_TOKENS (MT)	2
IGNORE	[“i”, “im”, “me”, “mine”, “you”, “yours”]
MIN_TWEETS_IN_SLICE (MTIS)	1
MIN_COSSIM_FOR_EXISTING (MCFE)	0.5
MIN_ENTROPY (ME)	3.5
MIN_UNIQUES (MU)	2
WINDOW_SIZE_IN_SECONDS (WSIS)	900s

to indicate that news tweets have higher entropy than ordinary tweets, which seem reasonable. Even though the value is set higher, there was not a huge difference in the paper's results between their suggested value and the value we found to give the highest precision for detecting news tweets.

Another parameter for which previous research seemingly has found an optimal value is the one for excluding tweets containing less than a certain number of tokens (*MIN_TOKENS* - MT). Both increasing and decreasing this value led to a slight decrease in precision.

One parameter we thought ahead of time would be important in detecting news is the *MIN_TWEETS_IN_SLICE* (MTIS) parameter. This is the number of tweets assigned to a cluster during the current time window. Although it seemingly has little impact on the detection of breaking news, this could change if tested on a data set that is recorded when some catastrophe or similar crisis happens. This is because the parameters work as sort of a dial; the higher it is set, the more tweets must be generated in a cluster during a window for it to be detected, which indicates the importance of the event the cluster describes.

The parameter for excluding clusters with less than a certain amount of unique users posting to it (*MIN_UNIQUES* - MU) sees little change in precision by increasing it by one from the baseline. Although precision increases slightly, by looking at the raw number one can see that the number of relevant clusters and the number of not relevant clusters decrease almost uniformly. Keeping it low, but still above 1 to filter out an ocean of clusters containing only singular messages, seem like the best course of action.

Finally, we have the parameter governing the size of the time window we examine (*WINDOW_SIZE_IN_SECONDS* - *WSIS*). This parameter, like MTIS, did not greatly affect the results. Still it demonstrated that unlike what we had assumed in advance, having a shorter time window did not in general impact the precision negatively. As we want to detect breaking news as soon as possible we want this value to be as low as possible.

The test labeled *FINAL* in Table II is thus a test where we have combined the optimal values for ME, MCFE, WSIS in addition to filtering the clusters using NER.

VI. DISCUSSION AND CONCLUSION

In this paper, we have looked at how one can use an algorithm intended for clustering documents (locality-sensitive hashing) to filter tweets. We have shown that, by tweaking the parameters of LSH and including a named-entity recognition engine in the pipeline, it is possible to achieve a high precision value for news events in tweets.

The LSH implementation itself performs well, and its memory footprint only grows linearly with the size of the vocabulary of the input. With the rate of the free Twitter streaming API, the implementation processes the tweets quicker than they arrive, meaning the implementation is usable in a real-time system.

Many of the relevant tweet clusters found by our algorithm were tweets and retweets from the U.S. Geological Survey about small earthquakes occurring in the San Francisco area (where our tweets were gathered from). These, to us, are indeed interesting events, however they would only likely be picked up in areas with seismic activity. Thus, our findings should be tested on data sets gathered from other geographic areas.

News detection from Twitter is important in the SmartMedia project[18], in which semantic technologies and various recommendation strategies are combined into a mobile news aggregator[19]. When a new event is detected from Twitter, the relevant tweets are fed into the news index and subjected to the same personalization approach as traditional news stories[20]. Whether these tweets provide new insight or faster insight into breaking news will be tested in future experiments.

ACKNOWLEDGEMENTS

This work is a part of the SmartMedia[7] Program in NTNU, Norway. The author Özlem Özgöbek is supported by the ERCIM Alain Bensoussan Fellowship Programme.

REFERENCES

[1] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, “Microblogging during two natural hazards events: What twitter may contribute to situational awareness,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’10. New York, NY, USA: ACM, 2010, pp. 1079–1088. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753486

[7]http://research.idi.ntnu.no/SmartMedia

ISSN 2395-8618

TABLE II
TEST WITH PRECISION MEASURE FOR THE DIFFERENT TESTS WHERE FINAL IS THE BEST LSH TUNING COMBINED WITH NER.

Test	Parameters	Relevant	Not Relevant	Precision
BASELINE	All at standard (see table I)	199	133	0.472
BASELINE + NER	All at standard	106	49	0.683
MIN_TOKENS #1	MT = 1	117	166	0.413
MIN_TOKENS #2	MT = 3	118	143	0.452
MIN_TWEETS_IN_SLICE #1	MTIS = 2	79	55	0.589
MIN_TWEETS_IN_SLICE #2	MTIS = 3	32	25	0.561
MIN_TWEETS_IN_SLICE #3	MTIS = 4	13	17	0.433
MIN_TWEETS_IN_SLICE #4	MTIS = 5	4	9	0.307
MIN_COSSIM_FOR_EXISTING #1	MCFE = 0.4	137	591	0.188
MIN_COSSIM_FOR_EXISTING #1 + NER	MCFE = 0.4	118	123	0.489
MIN_COSSIM_FOR_EXISTING #2	MCFE = 0.6	86	45	0.656
MIN_COSSIM_FOR_EXISTING #2 + NER	MCFE = 0.6	80	29	0.733
MIN_ENTROPY #1	ME = 3	121	345	0.259
MIN_ENTROPY #1 + NER	ME = 3	28	33	0.259
MIN_ENTROPY #2	ME = 4	53	23	0.697
MIN_ENTROPY #2 + NER	ME = 4	51	11	0.697
MIN_UNIQUES #1	MU = 1	55	56	0.495
MIN_UNIQUES #2	MU = 3	18	31	0.367
WINDOW_SIZE #1	WSIS = 300	107	78	0.578
WINDOW_SIZE #1 + NER	WSIS = 300	98	34	0.742
WINDOW_SIZE #2	WSIS = 600	116	114	0.504
WINDOW_SIZE #3	WSIS = 1200	114	144	0.441
WINDOW_SIZE #4	WSIS = 1500	112	150	0.427
FINAL	MT = 2, MTIS = 1, MCFE = 0.6, ME = 4, MU = 2, WSIS = 300	22	2	0.916

[2] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[3] S. Phuvipadawat and T. Murata, "Breaking news detection and tracking in twitter," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 3. IEEE, 2010, pp. 120–123.

[4] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma, "Breaking news on twitter," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 2751–2754.

[5] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 181–189.

[6] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[7] B. W. Locke, "Named entity recognition: Adapting to microblogging," 2009.

[8] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 359–367. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002519

[9] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan, "Domain adaptation of rule-based annotators for named-entity recognition tasks," ser. EMNLP '10, 2010, pp. 1002–1012.

[10] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, "Twiner: Named entity recognition in targeted twitter stream," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '12. New York, NY, USA: ACM, 2012, pp. 721–730. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348380

[11] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1524–1534. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145595

[12] P. Agarwal, "Prediction of trends in online social netwok," Ph.D. dissertation, Indian Institute of Technology New Delhi, 2013.

[13] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, 1998, pp. 604–613.

[14] C. E. Shannon, "A note on the concept of entropy," *Bell System Tech. J*, vol. 27, pp. 379–423, 1948.

[15] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[16] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 473–480.

[17] M. Vogiatzis, "Using storm for real-time first story detection," Master's thesis, University of Edinburgh, School of Informatics, 11 Crichton Street, Edinburgh, Midlothian EH8 9LE, Great Britain, 2012, retrieved online on May 14 2015 at url=https://micvog.files.wordpress.com/2013/06/vogiatzis_storm.pdf.

[18] M. Tavakolifard, K. C. Almeroth, and J. A. Gulla, "Does social contact matter?: modelling the hidden web of trust underlying twitter," in *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013, pp. 981–988.

[19] J. E. Ingvaldsen, J. A. Gulla, and Ö. Özgöbek, "User controlled news recommendations," in *Proceedings of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2015)*, 2015.

[20] J. A. Gulla, A. D. Fidjestøl, X. Su, and H. Castejon, "Implicit user profiling in news recommender systems," 2014.

ISSN 2395-8618

Optimize Hierarchical Softmax with Word Similarity Knowledge

Zhixuan Yang, Chong Ruan, Caihua Li, and Junfeng Hu

***Abstract*—Hierarchical softmax is widely used to accelerate the training speed of neural language models and word embedding models. Traditionally, people believed that the hierarchical tree of words should be organized by the semantic meaning of words. However, Mikolov et al. showed that high quality word embeddings can also be trained by simply using the Huffman tree of words. To our knowledge, no work gives a theoretic analysis on how we should organize the hierarchical tree. In this paper, we try to answer this question theoretically by treating the tree structure as a parameter of the training objective function. As a result, we can show that the Huffman tree maximizes the (augmented) training function when word embeddings are random. Following this, we propose *SemHuff*, a new tree constructing scheme based on adjusting the Huffman tree with word similarity knowledge. Experiment results show that word embeddings trained with optimized hierarchical tree can give better results in various tasks.**

***Index Terms*—Hierarchical Softmax, Word Embedding, Word Similarity Knowledge**

I. Introduction

Traditionally, words are treated as distinct symbols in NLP tasks. This treating has limitations especially when it is used with n-gram models. For example, if the size of the vocabulary is $|V|$, an n-gram language model will have $O(|V|^n)$ parameters. The curse of dimensions in the number of parameters leads to great difficulties on learning and smoothing the model. More advanced methods were proposed to address this problem. Word embedding, also known as distributed representations or word vectors, is among one of them. The key idea is to exploit the similarity between words. Word embedding maps words to vectors of real numbers in a low dimensional space. Similar words are mapped to close vectors while dissimilar words are mapped to vectors with longer distance. For example, the word *cat* may be mapped to (0.8, 0.7, ⋯), and *dog* may be mapped to (0.75, 0.77, ⋯), while the word *Paris* may be mapped to (-0.5, -0.9, ⋯).

Bengio et al. [1] proposed a neural language model that uses word embeddings as input features of the language model. The model also treats word embeddings as unknown parameters and learns them from data just like other parameters in the neural network. Such neural network methods were further improved in both efficiency and accuracy by many researchers in the past decade such as the log-bilinear model by Mnih & Hinton [14] and the skip-gram model by Mikolov et al. [11]. More efficient training algorithms were also introduced, including hierarchical softmax by Morin and Bengio [16], NCE (noise contrastive estimation) [4] [13], and negative sampling by Mikolov et al. [10], which is a further simplification of NCE. Simpler models and more efficient training algorithms allow learning accurate word embeddings from large-scale corpus.

People used to believed that the hierarchical tree used in hierarchical softmax should organize words by the semantic meaning of words. For example, Morin & Bengio [16] extracted the tree from the IS-A taxonomy in WordNet [12] and Mnih & Hinton [15] constructed the tree by repeatedly clustering word embeddings into a binary tree. However, Mikolov et al. showed that high quality word embeddings could also be learned by simply using the Huffman tree. To our knowledge, no work has given a theoretical analysis on how we should organize the hierarchical tree. Inspired by the analysis performed by Levy and Goldberg [9], we try to answer this question by treating the tree structure as a parameter of the training objective function. Following this, we show that the Huffman tree can maximize the augmented objective function when word embeddings are random. We can also explain more clearly why semantic related words should be placed together in the hierarchical softmax tree. Following the theoretical analysis, we propose **SemHuff**, a new hierarchical softmax tree constructing scheme based on adjusting the Huffman tree by rearranging nodes in same level. Our experiments show that SemHuff can improve the Huffman tree in all tasks and hierarchical softmax can outperform negative sampling in some situations.

The rest of this paper is organized in the following way: Section 2 analyzes hierarchical softmax by treating the tree structure as a parameter. Section 3 proposes SemHuff, our new tree constructing scheme. Section 4 presents our experiment results, which compares the negative sampling and the hierarchical softmax with different tree constructing schemes.

Manuscript received June 25, 2016.

Junfeng Hu is the corresponding author. (phone: 86-10-62765835 ext 103; fax: 86-10-62765835 ext 101; e-mail: hujf@pku.edu.cn)

Zhixuan Yang and Caihua Li are with the School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871, P. R. China. (e-mail: {yangzx95, peterli}@pku.edu.cn).

Chong Ruan and Junfeng Hu are with Key Laboratory of Computational Linguistics, Ministry of Education, Institute of Computational Linguistics, School of Electronics Engineering and Computer Science, Peking University, P. R. China. (e-mail: {pkurc, hujf}@pku.edu.cn).

This work is supported by the National Natural Science Foundation of China. (grant No. M1321005, 61472017).

ISSN 2395-8618

In the end, section 5 concludes this paper.

II. Analysis

A. Why Huffman Tree Works?

The skip-gram model predicts the probability over all words given one word in its context. The training objective function is the log likelihood:

$$L = \sum_{w_i \in Corpus} \sum_{w_j \in ctx(w_i)} \log\, p(w_j \mid w_i)$$

A linear classifier with $|V|$ outputs is used to predict the conditional probability $p(w_j|w_i)$. In order to avoid an output layer with $|V|$ outputs, which is very computationally expensive, hierarchical softmax organizes all words as a binary tree. Then, the conditional probability $p(w_j|w_i)$ of w_j is decomposed as the product of probabilities of going from the root of the tree to the leaf node representing w_j. When skip-gram is used with hierarchical softmax, the objective function can be written as:

$$L = \sum_{w_i \in Corpus} \sum_{w_j \in ctx(w_i)} \sum_{n \in path(w_j)} \log\, p(d(w_j, n) \mid n, w_i)$$

where $path(w_j)$ is the set of internal nodes on the path from the root of the tree to the leaf node representing w_j. And d (abbreviation for *direction*) is defined as:

$$d(w_j, n) = \begin{cases} 0, & w_j \text{ is in the left subtree of } n \\ 1, & w_j \text{ is in the left subtree of } n \end{cases}$$

The above objective function sums over every word and its context words in the corpus. We can group the summation by (w_i, w_j) pairs so that the equation can be written as:

$$L = \sum_{w_i \in V} \sum_{w_j \in V} \#(w_i, w_j) \sum_{n \in path(w_j)} \log\, p(d(w_j, n) \mid n, w_i)$$

where $\#(w_i, w_j)$ is the total number of observed (w_i, w_j) pairs in the corpus. We can divide the equation by the total number of all observed pairs, so that the count can be treated as probability:

$$L = \sum_{w_i \in V} \sum_{w_j \in V} p(w_i, w_j) \sum_{n \in path(w_j)} \log\, p(d(w_j, n) \mid n, w_i) \quad (1)$$

Then we change the order of the summations:

$$L = \sum_{w_j \in V} p(w_j) \sum_{n \in path(w_j)} \sum_{w_i \in V} p(w_i \mid w_j) \log\, p(d(w_j, n) \mid n, w_i)$$

where $p(w_i|w_j) = p(w_i, w_j) / p(w_j)$ is the conditional probability of w_i occurs in the context of w_j. When hierarchical softmax is used with the skip-gram model, $p(d(w_j, n)|\, n, w_i)$ is computed by a logistic regression:

$$p(d(w_j, n) \mid n, w_i) = \begin{cases} \sigma(\vec{n}^T \vec{w}_i), & \text{for } d(w_j, n) = 1 \\ 1 - \sigma(\vec{n}^T \vec{w}_i), & \text{for } d(w_j, n) = 0 \end{cases}$$

If word embeddings and classifier weights are independently random initialized from a uniform distribution with zero mean, the expectation of $p(d(w_j, n)|\, n, w_i)$ is 1/2. And the expectation of L with respect to word embeddings is:

$$E[L] = \sum_{w_j \in V} p(w_j) \sum_{n \in path(w_j)} \sum_{w_i \in V} p(w_i \mid w_j)\,(-1)$$

Since $p(w_i|w_j)$ is a distribution, it sums to 1:

$$E[L] = \sum_{w_j \in V} p(w_j) \sum_{n \in path(w_j)} (-1) = -\sum_{w_j \in V} p(w_j) \mid path(w_j) \mid$$

This is also the objective function of the Huffman tree.

From this result, we can see why the idea of using Huffman tree as the hierarchical softmax tree works in practice, since it maximizes the objective function when word embeddings are random. It also points out the structure of the hierarchical tree should consider the frequency of the words besides the meaning of the words.

B. Why Similar Words Should Be Placed Together?

The motivation of placing similar words in nearby positions in the tree is straightforward: the training samples for the classifier in each internal node will be more separable because similar words will have similar contexts by the distributional hypothesis of Harris [5]. We can explain this idea more clearly from the perspective of the training objective functions. It will also show some connections between the hierarchical softmax and decision trees.

In the training objective function (1), we can sums over all internal nodes first:

$$L = \sum_{n} \sum_{w_i, w_j \in V} p(w_i, w_j) \log\, p(d(w_j, n) \mid n, w_i)$$

If we assume the classifier in node n is perfectly trained, its estimation of the probability of going to the left subtree when the input word is w_i should be the ratio of training samples leading w_i to left, that is:

$$p(d = 0 \mid n, w_i) = \sum_{\substack{w_j \\ s.t.\ d(w_j, n) = 0}} p(w_j \mid w_i)$$

Hence an upper bound of the log-likelihood is:

$$L \le \sum_{n} \sum_{w_i} \sum_{k \in \{0,1\}} p(d{=}k, w_i|n) \log\, p(d{=}k|w_i, n) = -\sum_{n} \mathrm{H}_n[d|W]$$

where $\mathrm{H}_n[d \mid W]$ is the conditional entropy between the direction d and the word embedding W treated as a random variable, at internal node n. If we adopt a top-down splitting tree constructing scheme like the decision tree, the dividing criteria should be the conditional entropy. If we want to minimize the conditional entropy, placing words with similar contexts in the same subtree is likely to reduce the entropy.

III. SemHuff

In this section, we propose our hierarchical tree constructing scheme: *SemHuff*. Noticed that if we rearrange the nodes or subtrees of a Huffman tree within the same level, the resulting tree remains a Huffman tree, since all leaves keep their original depths unchanged. So our strategy is to rearrange nodes in the same level so that similar words can be placed together in the Huffman tree.

A. Description

Assuming we already have word similarity knowledge: S_{ij} is the similarity between word w_i and w_j. We generate an arbitrary Huffman tree from the corpus first. Then we adopt a bottom-up adjusting strategy: (1) from the bottom level to the top level, for level i, we calculate the similarity between all subtrees. The

ISSN 2395-8618

similarity of two subtrees is defined as the average similarity between theirs leaves (i.e. words). (2) Then we apply the weighted maximum matching algorithm to the similarity graph of subtrees. (3) Now, we rearrange these subtrees in level *i* according to the matching results. Matched subtrees are organized as siblings and subtrees without matching are paired arbitrarily. (4) go back to step (1) for level *i*-1 The algorithm is described by pseudo code in Figure 1.

```
function SemHuff(Corpus, S)
    T ← GenerateHuffmanTree(Corpus)
    for d from Depth(T) down to 1 do
        N ← {s | subtree s in level d of T}
        for i in N and j in N do
            S^d_{i,j} ← Average({S_{v,v'} | w_v ∈ i and w_{v'} ∈ j})
        end for
        matching ← WeightedMaximumMatching(N, S^d)
        for i ∈ N do
            if i is matched then
                rearrange i and its match as siblings in level d of T
            else
                randomly pick another unmatched subtree as its sibling
            end if
        end for
    end for
    return T
end function
```

Fig. 1. The pseudo code of SemHuff. The function takes the corpus and a similarity measure S between words as input and returns the adjusted Huffman tree.

The word similarity knowledge can be extracted from different kinds of language resources like WordNet [12] and PPDB [2] for the English language. For the Chinese language, we extract the similarity knowledge from the ontology generated by He et al. [6] in our experiments. The word similarity knowledge used by *SemHuff* is not limited to use external prior knowledge. It is also possible to use the word embeddings being trained by a bootstrapping process similar to Mnih and Hinton [15].

B. Demonstration

To illustrate our algorithm more clearly, a hand-crafted demo is provided for further investigation. We choose only 6 words and set the corresponding similarity matrix intuitively. Then the adjusting procedure is presented in Figure 2.

In the first step, we consider all possible matches among the leaves, *cat*, *dog*, *tree*, *grass* and find out the best match is obviously ⟨*cat*, *dog*⟩ and ⟨*tree*, *grass*⟩, so *cat* and *dog* are set as siblings and so are *tree* and *grass*. When it comes to the penultimate layer, the four nodes/subtrees to be matched are *plant*, {*dog*, *cat*}, *animal*, {*tree*, *grass*}. In this layer, the best matches are ⟨*animal*, {*dog*, *cat*}⟩ and ⟨*plant*, {*tree*, *grass*}⟩. Now we see that all similar words are grouped together. After this step, adjusting procedure stops, because no upper layer has more than two nodes.

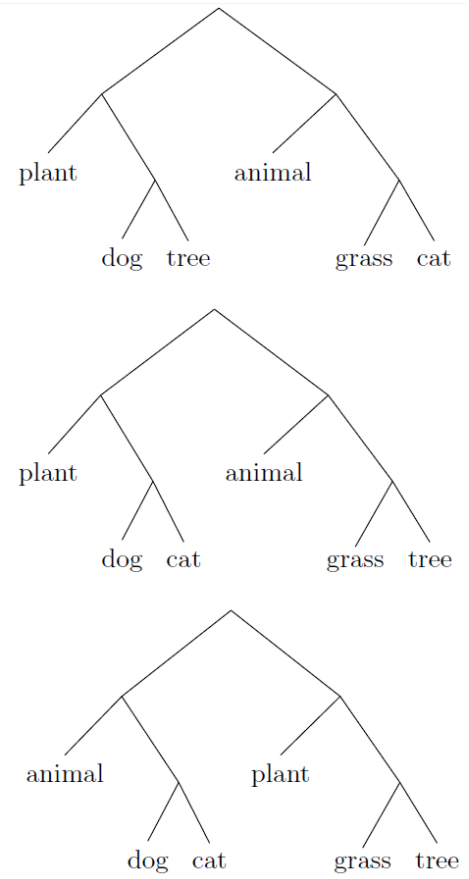

Fig. 2. Top: Huffman tree generated from corpus. Middle: Huffman tree after adjusted the bottom layer. Bottom: the final Huffman tree.

C. Implementation Details

Our implementation of *SemHuff* is modified from the original *word2vec* package. In our program, we use the highly efficient *Blossom V* package written by Kolmogorov & Vladimir [7] to perform weighted maximum matching. However, it's not practical to find the exact weighted maximum matching of a dense graph which has tens of thousands vertices and about a billion edges, since the complexity of the matching algorithm is $O(|E||V|^3)$. As a result, some approximation is made: for each word, we only keep its 30 most similar words and thus speed up the computation. With this method, adjusting can be done in less than 20 minutes for a Huffman tree of 80k leaves and 24 layers on a typical workstation.

We'd like to show a snippet of the adjusted Huffman tree in Figure 3. We see that similar words have been put together and it demonstrates the effectiveness of our adjusting algorithm.

IV. Experiments

We conduct experiments to compare the performance of different hierarchical tree structures. Following Lai et al. [8], we evaluate word embeddings by tasks from two perspectives: semantic properties of word embeddings and using word embeddings as features for downstream NLP applications.

Our experiments were conducted on the Chinese language.

ISSN 2395-8618

The training corpus used is Xinhua News [3] of two years: 1997 and 2004. There are about 50 million tokens in the corpus and 80 thousand words occurring at least 5 times.

```
<word>伊拉克(Iraq)</word>
<level10>
  <word>她(She)</word>
  <level11>
    <word>颗粒物(Particles)</word>
    <level12>
      <word>人权(Human Rights)</word>
      <level13>
        <level14>
          <word>稳(Stable)</word>
          <level15>
            <word>04</word>
            <word>董事(Director)</word>
          </level15>
        </level14>
```

```
<level8>
  <level9>
    <word>伊拉克(Iraq)</word>
    <level10>
      <word>俄罗斯(Russia)</word>
      <level11>
        <word>伊朗(Iran)</word>
        <level12>
          <word>阿富汗(Afghanistan)</word>
          <level13>
            <word>利比亚(Libya)</word>
            <word>波黑(Bosnia and Herzegovina)</word>
          </level13>
        </level12>
      </level11>
```

Fig. 3. Top: Huffman tree generated from corpus. Bottom: Huffman tree after adjustment.

The following models are compared in our experiments:

a) HS-Huffman: hierarchical softmax with Huffman tree.

b) HS-SemHuff: hierarchical softmax with SemHuff. The similarity knowledge used is the ontology generated by the hierarchical clustering algorithm by He et al. [6].

c) NS: negative sampling with 7 negative samples.

All models are used with skip-gram and run 20 iterations over the entire corpus. The subsampling rate is set to 10^{-5}.

A. *Word Similarity Task*

We extract 3020 pairs of synonyms from *Tongyici Cilin* (available at http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm), which is a manually built Chinese thesaurus. *Tongyici Cilin* comprises sets of Chinese synonyms. The synonym pairs are chosen from synonym sets whose size is not greater than 10, because large synonym set in *Tongyici Cilin* tends to be inconsistent.

For each extracted synonym pair 〈A, B〉, we measure the rank of B in a set of candidate words by the distance of its word embedding to the word embedding of A. The mean rank (MR) and mean reciprocal rank (MRR) is used to evaluate the quality of learned word embeddings. For MR, lower is better; while for MRR, higher is better. The result for different models in different settings is showed in Figure 4 and Figure 5.

The MR and MRR give consistent evaluation: NS > HS-SemHuff > HS-Huffman. Though, the MR value of several hundred is somehow counterintuitive. After investigation, the reason can be explained as follows. Only one word embedding is learned for a certain word, while it is always the case that one word has multiple meanings and usages. Take 团长 and 参谋长 for an example, 参谋长 means "a chief of staff in an army", while 团长 have many meanings. One of its meanings is "a regimental commander", which is used in the context of military affairs, so it may co-occur with 参谋长 frequently. Besides, 团长 can also be used to refer to the head of a delegation, a circus troupe, an opera troupe, and many other groups. This

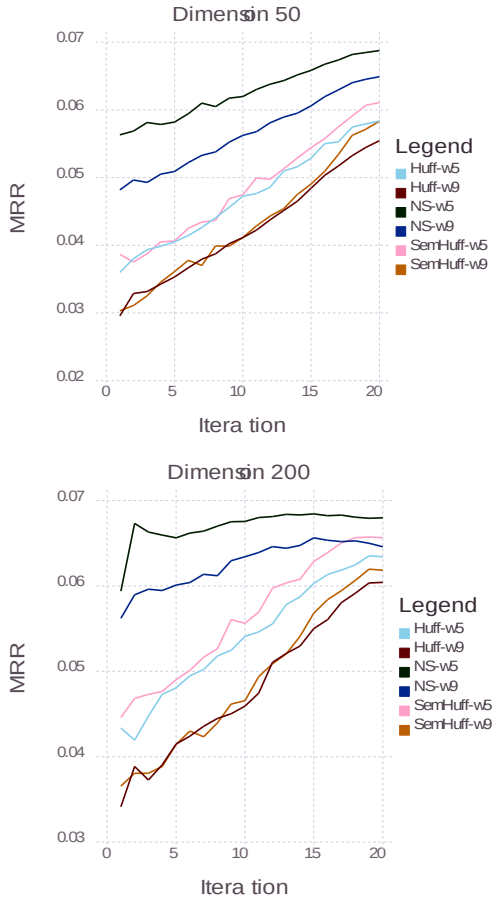

Fig. 4. Word similarity test results measured by MRR(mean reciprocal rank). The x-axis is the number of iterations over corpus during the training of the word embedding. The first row presents results for 50-dimensional vectors, while the second row is for 200-dimensional vectors.

asymmetry in word usage leads to the following phenomenon: for the word 参谋长, 团长 is its close neighbor, because 参谋长 is only used in military topics and 团长 is also salient in this context. But when we stand in the view of 团长, 参谋长 is not similar to 团长, because many other words such as 演员(actor), 代表(representative) will occupy the vicinity of 团长, and 参谋长 will get a rank worse than 1000. If we average a small number and a large number, say 10 and 1000, the result will be several hundred. What's more, the training corpus and the testing word pairs are not exactly in-domain, which contributes to this large rank, too.

B. *Analogy Task*

Our second evaluation method is the analogy task. If the relation between word A and B is similar to the relation between C and D, then *vector*(A) - *vector*(B) ≈ *vector*(C) - *vector*(D). Since we use Xinhua News as training data, information about Chinese cities and provinces should be attained. Thus, we choose 20 pairs of 〈provincial capital city〉:〈province〉 to

ISSN 2395-8618

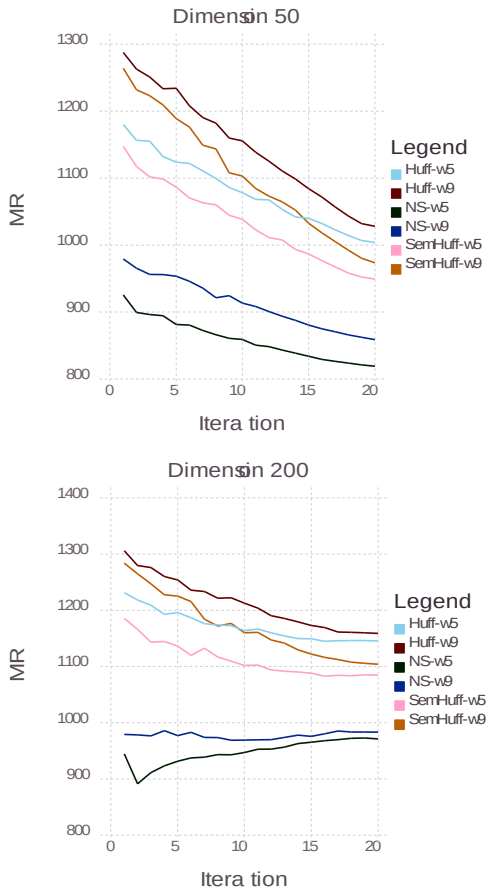

Fig. 5. Word similarity test results measured by MR(mean rank). The x-axis is the number of iterations over corpus during the training of the word embedding. The first row presents results for 50-dimensional vectors, while the second row is for 200-dimensional vectors.

generate $20^2 = 400$ analogy problems. In each problem, a city-province pair, say A and A', and another city, say B, are given, while the province whose capital city is B is unknown. Then, we search the word X from the whole vocabulary to fill in the blank such that the analogy identity *vector*(city A) – *vector*(province A') = *vector*(city B) – *vector*(X) is fitted as well as possible.

All models were trained for 20 iterations and the test is performed after every iteration. The best result for each model is showed in Table 1. The result is encouraging: word embeddings always give some province as the answer although we search for the word X among all words in the vocabulary. A precision of 75% can be achieved with our *SemHuff* model.

From these results, we see that NS is powerful for dimension 50 but is surpassed by hierarchical softmax models in the high dimensional case. Our *HS-SemHuff* model improves HS-Huffman significantly in the 50-dimensional case, and it outperforms HS-Huffman in every experimental setting, and gives the best result for dimension 200.

C. *POS Tagging Task*

The third task is using word embeddings as features for POS tagging. In this task, we use a very simple POS tagger: for a word *w*, we concatenate the word embeddings of words in the context windows of *w* as features.

TABLE I
ANALOGY TEST RESULTS

Model Name	Context Window Size	Word Vector Dimension	# of Correct Predictions
Huff	5	50	143
NS	5	50	**235**
SemHuff	5	50	217
Huff	9	50	161
NS	9	50	**254**
SemHuff	9	50	220
Huff	5	200	267
NS	5	200	194
SemHuff	5	200	**280**
Huff	9	200	294
NS	9	200	268
SemHuff	9	200	**300**

And a linear softmax classifier is trained on these features. This simple model is used because we think simpler models can better reflect the quality of input features. The results are showed in Figure 6.

In this task, NS performs better. As for two hierarchical softmax models, our *HS-SemHuff* is comparable with HS-Huffman in the low dimensional case and gives better results for dimension 200.

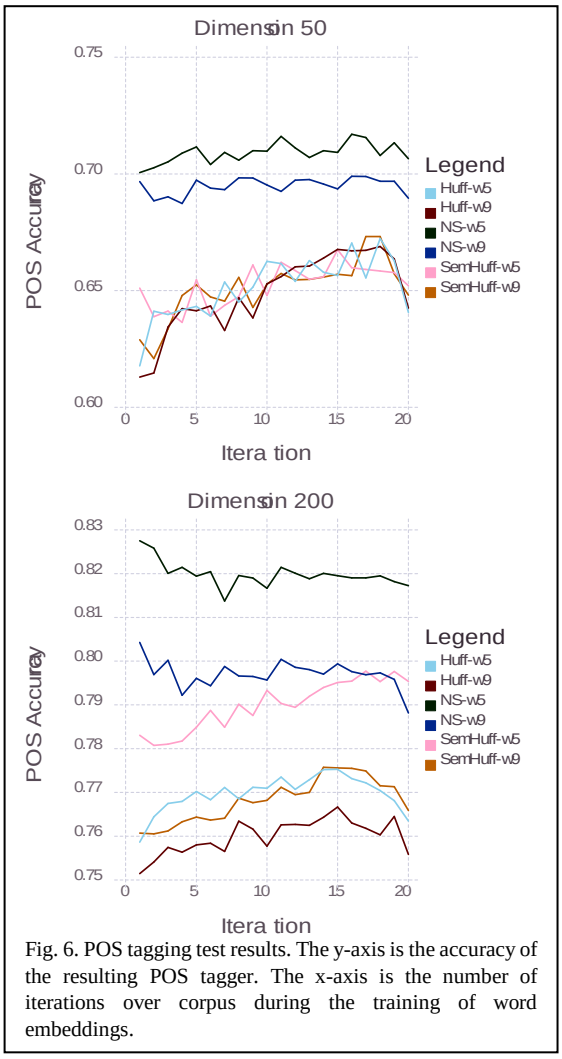

Fig. 6. POS tagging test results. The y-axis is the accuracy of the resulting POS tagger. The x-axis is the number of iterations over corpus during the training of word embeddings.

ISSN 2395-8618

V. Conclusion

In this paper, we seek to give some theoretical analysis on the widely used hierarchical softmax algorithm. By treating the tree structure as a parameter of the training objective function, we show that the reason why the common practice of using the Huffman tree works well is that the Huffman tree maximizes the objective function when word embeddings are random. We also show that the dividing criterion is the conditional entropy if we adopt a top-down splitting tree constructing scheme. Following the theoretical analysis, we propose *SemHuff*, a tree constructing scheme based on adjusting the Huffman tree. From the experiments, we show that negative sampling performs well in most tasks while hierarchical softmax performs better in high dimensional analogy task. And *SemHuff* further improves the original hierarchical softmax algorithm in all of our tasks.

In fact, a more natural idea is to directly optimize the training objective function with respect to the tree structure instead of adjusting a Huffman tree. However, the optimization over the space of all binary trees seems hard. We think some approximation or relaxation is necessary to solve this optimization problem. We leave this as future work for this research.

References

[1] Yoshua Bengio et al., "A Neural Probabilistic Language Model," in: *The Journal of Machine Learning Research* 3 (2003), pp. 1137 – 1155.

[2] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch, "PPDB: The Paraphrase Database," in: *Proc. of NAACL-HLT*, Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 758 – 764.

[3] David Graff and Ke Chen, "Chinese Gigaword," in: *LDC Catalog No.: LDC2003T09, ISBN* 1, 2005, pp. 58563 – 58230.

[4] M Gutmann and A Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in: *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 1 – 8.

[5] Zellig S Harris, "Distributional structure," in: *Word* 10.2-3 (1954), pp. 146 – 162.

[6] Shaoda He et al., "Construction of Diachronic Ontologies from People's Daily of Fifty Years," in: *Proc. of the Ninth International Conference on Language Resources and Evaluation*, Edited by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014.

[7] Vladimir Kolmogorov, "Blossom V: A new implementation of a minimum cost perfect matching algorithm," in: *Mathematical Programming Computation* 1.1, 2009, pp. 43 – 67.

[8] Siwei Lai et al., 2015. "How to Generate a Good Word Embedding?." Available: http://arxiv.org/abs/1507.05523

[9] Omer Levy and Yoav Goldberg, "Neural Word Embedding as Implicit Matrix Factorization," in: *NIPS*, 2014, pp. 1 – 9.

[10] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, "Linguistic regularities in continuous space word representations," in: *Proc. of NAACL-HLT*, June. 2013, pp. 746–751.

[11] Tomas Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," in: *Proc. of the International Conference on Learning Representations*, 2013, pp. 1–12.

[12] George A Miller, "WordNet: a lexical database for English," in: *Communications of the ACM 38.11*, 1995, pp. 39–41.

[13] Andriy Mnih, "Learning word embeddings efficiently with noise-contrastive estimation," in: *NIPS*, 2013, pp. 1–9.

[14] Andriy Mnih and Geoffrey Hinton, "Three new graphical models for statistical language modelling," in: *Proc. of the 24th international conference on Machine learning*, 2007, pp. 641–648.

[15] Andriy Mnih and Geoffrey E Hinton, "A Scalable Hierarchical Distributed Language Model," in: *NIPS*, 2008, pp. 1081–1088.

[16] Frederic Morin and Y Bengio, "Hierarchical probabilistic neural network language model," in: *Proc. of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 246–252.

ISSN 2395-8618

Comparing Sentiment Analysis Models to Classify Attitudes of Political Comments on Facebook (November 2016)

Chaya Liebeskind, Karine Nahon, Yaakov HaCohen-Kerner, Yotam Manor

***Abstract*—This paper is a preliminary study which compares nine ML methods of sentiment analysis aimed towards classifying a corpus of 5.3 million messages of the public on Facebook pages of incumbent politicians. Two sentiments were examined: the general attitude of a comment and the attitude of the comment towards the content of a political post. Our results show that Logistic Regression outperformed the other eight ML models in terms of accuracy and F-measures. Also, we found that n-gram representation performed best. An interesting finding is a difference in success rate when classifying attitude in general vs. attitude towards the content in the political context.**

***Index Terms*— Machine Learning, Political discourse, Sentiment analysis, , Social media**

I. Introduction

RESEARCH about the use of social media platforms, such as Facebook and Twitter, by politicians has increased in recent years. These studies examined patterns of behavior of politicians, characteristics of the relationships between politicians and other groups like journalists, celebs and influencers, success and failure factors of political use, propagation of political information in social media and more. This paper adds to the rich literature of politicians and social media by comparing nine Machine Learning (ML) methods of sentiment analysis in an attempt to classify a large corpus of 5.3 Million posts of users replying to politicians (Israeli Member of Knesset, hereafter MKs), posted on Facebook during 2014-2015. This is the first phase of a larger project aimed towards establishing an explanatory model for commenting positively on politicians posts on facebook. The goal of this first phase is to choose the best method for classifying automatically such a big corpus of comments on political posts, in order to be able later run statistical tests to develop an explanatory model of such comments.

In this research, we adopt a supervised ML approach. First, we obtained a user comments dataset annotated with sentiment. We distinguish between two sentiment classification tasks: General attitude and Attitude towards the content of the post. Second, we represent each comment as a vector of features. Our feature set include both Facebook depended features, such as "like" and emojis counts, and text-based features. We compare five different text representation approaches, i.e., word, lemma, character n-grams, dictionary-based and extended dictionary-based, by training a classifier to distinguish among sentiment labels, analyzing the relevant features and predicting sentiments for new comments.

The contribution of this study is derived by several factors: the dataset is derived from a large corpus (~5.3 Million messages posted over 2 years on Facebook), the comparison of two different sentiment classification tasks, and it is the first work in NLP on Hebrew Facebook for classification purposes.

II. Theoretical Background and Literature

A. Politicians on Social Media

Social media has an important impact on public discourse, and is a major player in political context by users and politicians. Comparative literature survey shows that the use of social media among politicians is constantly increasing in democracies, such as Britain [1], New Zealand [2], Australia [3], the US [4] and Israel [5], while also political participation on social media has increased. In the context of our study two main streams of research which examine political discourse on social media are relevant. One, research that focuses on information flows around political content, and on analysis of relationships among users. For example, Kushin and Kitchener focused on political groups on Facebook and found that the representation of viewpoints was highly skewed in favor of discussion among likeminded participlants (homophily) [6]. This homophilous tendency has been reported in other studies which examined other platforms such as Twitter and blogs [7, 8]. Second, research that focuses on sentiment in context of political discourse. For example, Robertson et al studied political discourse on Facebook while focusing on two politicians for 22 months and found that positive comments decreased over time, while negative comments increased [9]. This is similar to the findings of other researchers who showed that the political discourse is dominated by a small portion of users and has a large negative rhetoric laced with sarcasm and humor [10], and that online political discussion tends to

1 Manuscript received August 12, 2016.

Chaya Liebeskind is with the Jerusalem College of Technology, Lev Academic Center, Israel (e-mail: liebchaya@gmail.com).

Karine Nahon is with the Interdisciplinary Center Herzliya, Israel and University of Washington (e-mail: karineb@uw.edu)

Yakkov HaCohen-Kerner is with the Jerusalem College of Technology, Lev Academic Center (e-mail: kerner@jct.ac.il).

Yotam Manor is with the Hebrew University(e-mail: yotammanor@gmail.com).

ISSN 2395-8618

contain a significant level of uncivil discussion [6]. Stieglitz and Dang-Xuan have shown that emotionally charged Twitter messages tend to be shared more often and more quickly compared to neutral ones [11]. Our project enters at this domain. It contributes to the literature by examining the comments of the public on a large corpus of data (5.3 Million messages) collected for two years on posts of political incumbent in Israel (MKs).

B. Sentiment Analysis

When Automatic sentiment analysis addresses the tasks of automatically identifying, extracting, and analyzing subjective information in natural language texts. The general aim is to determine the author's opinion about a specific topic. Most sentiment analysis studies address marketing and commercial tasks, such as extracting opinions from customer reviews [12–14], movie reviews [15, 16], and product reviews [17, 18].

Simultaneously, there is increasing interest in the sentiment analysis of the social web. Sentiment analysis enables to know what people think about specific topic and to perform analysis in order to plan future actions. There is a widespread variety of studies concerning sentiment analysis of posts in various social forums such as: blogs, Facebook, and Twitter.

Tsytsarau and Palpanas [19] reviewed the development of sentiment analysis and opinion mining during the last years, and also discussed the evolution of a relatively new research direction, namely, contradiction analysis. The authors supplied an overview of the most popular sentiment extraction algorithms, used in subjectivity analysis and to compare between them. They also introduced an overview of the most popular opinion mining datasets and data sources. According to their analysis, the trends of the past years show an increasing involvement of the research community, along with a drive towards more sophisticated and powerful algorithms. They tried to identify several interesting open problems, and to indicate several promising directions for future research.

Various general approaches have been proposed for the sentiment classification task. Two of the main approaches are the ML and the Dictionary approaches. In our study, we used both the ML and the Dictionary approaches.

The ML approach is composed of two general steps: (1) learn the model from a training corpus, and (2) classify a test corpus based on the trained model [17, 20, 21]. Various ML methods have been applied for sentiment classification. For instance, Pang and Lee applied three ML methods: Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM) [22]. Pang and Lee [22] combined SVM and regression (SVR) modes, with metric labelling. Glorot et al. [23] applied a deep learning method for large-scale sentiment classification. Moraes et al. [24] empirically compared between SVM and ANN for document-level sentiment classification.

The Dictionary approach is based on a pre-generated dictionary that contains sentiment polarities of single words, such as the Dictionary of Affect of Language[2], the General Inquirer[3], the WordNet-Affect[4], or the SentiWordNet [25]. Polarity of a sentence or document is usually computed by averaging the polarities of individual words. Most of the dictionary methods aggregate the polarity values for a sentence or document, and compute the resulting sentiment using simple rule-based algorithms [26]. More advanced systems, such as the Sentiment Analyzer introduced by Yi et al. [21], and the Linguistic Approach by Thet et al [16], extract sentiments precisely for some target topics using advanced methods that exploit domain-specific features, as well as opinion sentence patterns and Part-Of-Speech tags.

Some studies applied both the ML and the Dictionary approaches. For example, Ortigosa et al. [27] introduced their system, called SentBuk, which is able to extract information about the student's sentiments from the messages they write in Facebook with high accuracy. SentBuk retrieves messages written by users in Facebook and classifies them according to their polarity (positive, neutral or negative), extracts information about the users' sentiment polarity according to the sent messages, models the users' regular sentiment polarity, and detects significant emotional changes. The classification method implemented in SentBuk combines lexical-based and ML methods. SentBuk obtained an accuracy result of 83.27% using this classification method. Thelwall, et al. [28] described and assessed the SentiStrength 2 as a general sentiment strength detection algorithm for the social web. Their software primarily uses direct indications of sentiment. The results from six diverse social web data sets (MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums) indicate that their software is better than a baseline approach for all data sets in both supervised and unsupervised cases. SentiStrength 2 is not always better than ML approaches that exploit indirect indicators of sentiment, and is particularly weaker for positive sentiment in news-related discussions. In general, SentiStrength 2 is robust enough to be applied to a wide variety of different social web contexts.

III. METHODS

We compare nine ML methods on a manually coded dataset (N=577) in order to find the best suitable algorithm for classifying comments in political pages of incumbent politicians on Facebook. Once we find the best method we can then classify automatically the entire corpus. The corpus is comprised of posts of 84 (out of 120) MK members (n posts = 33,537), and the comments of ~2.9M users (n of comments = ~5.3M).

A. Preparing the dataset for Analysis

We study two main variables: ATTITUDE and ATTITUDE_TOWARDS_CONTENT_OF_THE_POST.

ATTITUDE: The general attitude conveyed in a comment to a political message. The general attitude focuses on the vibe of the comment. For example – if a comment strengthens the

2 http://www.hdcus.com/

3 http://www.wjh.harvard.edu/~inquirer/

4 http://wndomains.fbk.eu/wnaffect.html

ISSN 2395-8618

opposite view of an MK post, this will still be considered as a general attitude that is positive.

ATTITUDE_TOWARDS_CONTENT_OF_THE_POST: The Attitude of the comment towards the political post denotes whether the commenter support or oppose the political content of the post (1=Positive, 2=Negative, 3=Neutral, 4= Not Applicable, that is the comment does not relate to the post of the MK, 99=Unclear/Undefined)

Initially, we manually coded 100 comments by 3 coders. Coding manually political messages is complicated as the same text may reflect multiple attitudes towards multiple stakeholders. Therefore, we needed 3 rounds of manual coding in order to reach a satisfactory reliability level. In each one of the rounds the coders discussed the disagreements and refined the coding scheme to reach a better agreement. In the 3rd round we calculated Fleiss' Kappa to measure reliability of agreement for two variables: attitude (0.78) and attitude towards content of the post (0.82). A Fleiss Kappa between 0.6-0.8 is considered a 'substantial agreement', and >0.8 'almost perfect agreement'[30]. Once we reached a high level of agreement, one coder continued and manully coded 612 comments. The comments were chosen respective to their distribution in the main corpus (see table 1).

TABLE I
THE DISTRIBUTION OF THE ATTITUDES IN THE SAMPLED DATASET

Variables	Positive	Negative	Neutral	Not Applicable	Unclear
ATTITUDE	233	327	47	-	5
ATTTUDE_TOWARDS_CONENT_OF_THE_POST	221	122	16	243	10

For the preparation of our dataset we omitted the unclear category and comments which were not written in Hebrew or in English. Finally, the dataset that we ran was N=577.

B. Supervised attitude classification

In this research, we adopt a supervised Machine Learning (ML) approach for classifying Facebook comments. We next describe the collected information from the text and Facebook properties and how we incorporate it as features within the ML framework.

Feature Sets. We next detail how the special characters of Facebook, e.g. emojis, found useful in prior work, are encoded as features and describe different text representations, which we have explored, for feature extraction.

Facebook-based Features. In the last decade, the necessity of incorporating Emojis' information in automated sentiment classification of informal texts was proven [14, 31–34]. Therefore, we encoded each Emoji as separate feature and counted the number of its occurrences in the comment. Next, using Facebook API, we extracted additional three Facebook depended features: the number of "likes" that the comment got, the number of comments on the comment, and a Boolean feature, which indicates whether the commentator also "liked" the status. Another two features that we defined are the number of occurrences of the MK writer of the post and the number of occurrences of other MKs, either aliens or rivals of the post writer.

Text-based Features. First, we define two general text-based features: the number of words in the comment and the number of characters in the comment. Then, following the rationale of Aisopos et al. [35] that the higher the number of punctuations is, the more likely is the corresponding comment to be subjective, we encoded common punctuations (with frequency > 10) as features by counting their normalized number of occurrences. In Twitter, Aisopos et al. found that while exclamation marks constitute a typical annotation for positive sentiments, question marks usually express a negative feeling. The defined punctuation features allow us to explore whether these findings are also valid in our setting.

Next, we investigate five types of text representations:

1. Unigram/Word representation - Each of the words in the comment is considered as a feature. The score of the feature is the word number of occurrences in the comment divided by the comment length (termed normalized word count).
2. Lemma representation- We lemmatized all the comments using a Part-of-Speech (PoS) tagger [36]. Then, each of the comments' lemmas is a feature scored by the normalized lemma count.
3. Character n-grams representation - Each comment is considered as a *character n-grams*, i.e., strings of length n. For example, the character 3-grams of the string "character" would be: "cha", "har", "ara", "rac", "act", "cte", and "ter". Since there is much less character combinations than word combinations, this representation overcomes the problem of sparse data that arises when using word representation. On the other hand, this representation still produces a considerably larger feature set. Previous work on short informal data showed that character n-gram features can be quite effective for sentiment analysis [35, 37]. This is due to the tendency of noise and misspellings to have smaller impact on substring patterns than on word patterns. Therefore, in this representation, we considered each of the character n-grams of the comment as a feature and scored it by its normalized count in the comment.
4. Dictionary-based representation - We combine the dictionary approach, which relies on a pre-built dictionary

ISSN 2395-8618

that contains opinion polarities of words, with our ML approach. Our features are the dictionary words scored by their normalized count. We used the intersection of the seed sentiment list with the manually extended list of 85 positive words and 83 negative words generated by HaCohen-Kerner and Badash [38].

5. Extended dictionary-based representation- We extended our dictionary with Facebook sentiment words by applying a statistical measure of word co-occurrence. Assuming that words that occur frequently together are topically related [39], for each sentiment word in the original dictionary (described in the previous dictionary-based representation), we extracted the 20 most similar word using *Dice coefficient* [40] and an unannotated corpus of over than 4 million comments. Then, an annotator selected the sentiment words from these candidate lists. We increased the size of our Hebrew dictionary (the extended sentiment list) from 177 words to 830 words (327 positive words and 503 negative words). Our features are the dictionary words scored by their normalized count. Since the dictionary was generated from the Facebook corpus, the extracted sentiment words are typical to Facebook. We recognized two interesting type of words: slang sentiment words such as ″king″ and ″stupid″, and sentiment words from events that affect political discourse such as ″terrorist attack″ and ″unemployed″.

IV. Resaults

We used nine ML methods to combine the features in a supervised classification framework: Random Forest, Decision Tree, Bagging, Adaboost, Bayes Network, Supported Vector Machine (SVM), Logistic Regression and Multilayered Perceptron. We estimated the accuracy rate of each ML method by a 10-fold cross-validation test. We ran these ML methods by the WEKA platform [41, 42] using the default parameters. To reduce the number of features in the feature sets, we tried to filter out non-relevant features using two well-known feature selection methods: Information gain (InfoGain, IG) [43] and Correlation-based Feature Subset (CFS) [44]. The second method had better performance. Therefore, the results presented in this section include CSF feature selection which significantly improved the accuracy of all the configurations. (We detail the important features, which were selected by the CSF feature selection for the best configurations in Table 5 of the analysis Section).

TABLE II
Comparison of results obtained by nine ML methods

#	ML Method	ATTITUDE		ATTITUDE_TOWARDS_CONTENT	
		Accuracy (%)	**F-Measure**	**Accuracy (%)**	**F-Measure**
1	Random Forest	74	0.713	60	0.589
2	Decision Tree (J48)	71	0.699	61	0.565
3	Bagging	73	0.712	63	0.598
4	AdaBoost (M1)	69	0.67	61	0.54
5	Bayes Network	71	0.693	60	0.55
6	Logistic Regression	**78**	**0.771**	**66**	**0.64**
7	Multilayered Pereceptron	75	0.744	62	0.595
8	SVM (SMO)	72	0.709	62	0.579
9	SVM (LibSVM)	69	0.673	61	0.54

TABLE III
Comparison of results obtained by five text-based representations

#	Representation	ATTITUDE		ATTITUDE_TOWARDS_CONTENT	
		Accuracy (%)	**F-Measure**	**Accuracy (%)**	**F-Measure**
1	Word\Unigram	78	0.771	66	0.64
2	Lemma	77	0.761	64	0.619
3	Character n-grams	**80**	**0.801**	**67**	**0.665**
4	Dictionary-based	74	0.733	61	0.554
5	Extended dictionary-based	75	0.742	62	0.562

Table 2 shows the performances of the different ML methods on the feature set of Facebook and the state-of-the-art word representation. The best ML method was Logistic Regression. Therefore, we have performed further experiments using only this method.

In this research, we investigated five types of text representations (Section 3): unigram\word representation, lemma representation, character n-grams representation, dictionary-based representation and extended dictionary-based representation. The attitude classification results of the Logistic Regression algorithm using each of these representations are presented in the left side of Table 3. The character n-grams representation (n=3) yielded the best accuracy result (80%). The advantage of the representation over the extended dictionary-based representation is notable (5%) and is statistically significant according to the McNamar

ISSN 2395-8618

test [45] at level 0.05. Even though, we extended our dictionary using statistical co-occurrence measure, the dictionary coverage is still limited. We consider utilizing a semi-automatic iterative scheme to increase the recall of the dictionary [46].

The results of the attitude towards content classification results of the Logistic Regression algorithm are presented in the right side of Table 3. The best results (67%) were obtained using the character n-grams representation (n=2 and n=3). However, these results are significantly lower than the results of the attitude classification. The task of attitude towards content classification is difficult and more sophisticated text understanding approaches, e.g. semantic similarity between the post and the comment, should be applied.

We experiment three configurations of the character n-grams representations: n=2, n=3 and a combination of n=2 and n=3. Table 4 shows a comparison of the character n-grams configurations for the two classification tasks. The optimal configurations of the tasks were different.

TABLE IV
A COMPARISON OF THE CHARACTER N-GRAMS CONFIGURATIONS

#	Character n-grams	ATTITUDE		ATTITUDE_TOWARDS_CONTENT	
		Accuracy (%)	**F-Measure**	**Accuracy (%)**	**F-Measure**
1	n=2	74	0.737	64	0.625
2	n=3	**80**	**0.801**	62	0.577
3	n=2 and n=3	74	0.75	**67**	**0.665**

V. ANALYSIS

We used the information obtained by the CFS selection method to better understand which features have more influence on the classification accuracy. Table 5 presents information on the features, which were selected by the CFS method for the best configuration f or each of the classification tasks. The Boolean feature, which indicates whether the commentator also "liked" the status was informative for both of the tasks. Although the name of the writer of the post was not selected by the attribute towards content classifier, the character 2-grams "MK", which indicates a mention of a politician, was selected. No emoji feature was selected for any of the tasks. Only some of the selected features were informative, namely formed a word of two or three letters in English or Hebrew. For example, in the top-20 selected features, both classification tasks selected the English word "age" along with the Hebrew words ″already″ and ″next″. Additional selected features for the attitude classification task were the Hebrew words ″law″ ,″father ,″ ″her″ ,″for them″ ,″past″ and ″white″. Additional selected features for the attitude towards content classification task were the Hebrew words ″sex″ ,″it″ ,″no″ and two plural suffixes of two letters.

TABLE V
SELECTED FEATURES FOR THE BEST CONFIGURATIONS

Task	# feat.	Facebook-based	Text-based	General
Attitude classification	62	COMMENTOR_LIKED, MK_WRITER_OF_POST	HE: 51 EN: 6	Special chars: "and ″ Comment length (number of words)
Attitude towards content	43	COMMENTOR_LIKED	HE: 20 (n=2), 14 EN: 3 (n=2), 4 (n=3)	Special chars: /

TABLE VI
ATTITUDE CLASSIFICATION: CONFUSION MATRIX

Positive	**Negative**	**Neutral**	
177	48	0	**Positive**
34	270	14	**Negative**
4	14	16	**Neutral**

TABLE VII
ATTITUDE TOWARDS CONTENT CLASSIFICATION: CONFUSION MATRIX

Positive	**Negative**	**Not applicable**	
129	69	19	**Positive**
26	212	13	**Negative**
30	31	48	**Not applicable**

ISSN 2395-8618

Table 5 we complete our analysis by presenting the confusion matrixes of the best classification results. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class.

Table 6 shows that most of the classification errors were due to incorrect classification of positive comments as negative (48) and vice versa (34). Most of the incorrectly classified neutral comments were classified as negative. No positive comment was classified as neutral.

Table 7 shows that most of the classification errors were due to incorrect classification of positive comments as negative (69) and vice versa (26). However, there was no different between the number of comments that are not forwarded to the content which were classified as positive (30) and the number of these comments which were classified as negative (31).

VI. Conclusions

In this paper, we presented two sentiment classification tasks: General attitude and Attitude towards the content of the post. We combined Facebook-based and text-based features in supervised ML algorithms. We obtained that classifying the attitude towards the content is significantly more difficult. For both of the tasks, we found that the character n-grams model text representation outperformed other four representations. This is the first work in NLP on Hebrew Facebook for classification purposes.

We further plan to explore word embedding for text reoresentations, where words are mapped to vectors of real numbers. Methods of word embedding mathematically reduce the dimension of the words' vector to a continuous vector with a lower dimension. The dimension reduction is often implemented by one of the following methods: neural networks, dimensionality reduction on the word co-occurrence matrix and probabilistic models.

In addition, to increase the performance of the attitude towards context classification, we plan to add features which calculate the textual and semantic similarities between the post and comment text.

ACKNOWLEDGMENT

We would like to express our deep gratitude to Avital Day, our research assistant, for her help in programming and carrying out the research experiments. This work was partially funded by an internal research grant from Jerusalem College of Technology, Lev Academic Center.

References

1. Williamson, A.: MPs on Facebook. Hansard Society, London (2009).
2. Busby, C., Bellamy, P.: New Zealand Parliamentarians and online social media. New Zealand Parliament's Parliamentary Library (2011).
3. Bruns, A., Burgess, J.: #Ausvotes: How twitter covered the 2010 Australian federal election, http://search.informit.com.au/documentSummary;dn=627330171744964;res=IELHSS.
4. Smith, A.: Cell Phones, Social Media and Campaign 2014, http://www.pewinternet.org/2014/11/03/cell-phones-social-media-and-campaign-2014/, (2014).
5. Haleva-Amir, S.: Personal Web Applications in the Service of Knesset Members: Personal Israeli Politics in the DIgital Era, (2014).
6. Kushin, M.J., Kitchener, K.: Getting political on social network sites: Exploring online political discourse on Facebook. First Monday. 14, (2009).
7. Nahon, K., Hemsley, J.: Homophily in the Guise of Cross-Linking Political Blogs and Content. Am. Behav. Sci. 0002764214527090 (2014).
8. Colleoni, E., Rozza, A., Arvidsson, A.: Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. J. Commun. 64, 317–332 (2014).
9. Robertson, S.P., Douglas, S., Maruyama, M., Semaan, B.: Political Discourse on Social Networking Sites: Sentiment, In-group/Out-group Orientation and Rationality. Inf Polity. 18, 107–126 (2013).
10. Mejova, Y., Srinivasan, P., Boynton, B.: GOP Primary Season on Twitter: "Popular" Political Sentiment in Social Media. In: WSDM13 (2013).
11. Stieglitz, S., Dang-Xuan, L.: Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. J. Manag. Inf. Syst. 29, 217–248 (2013).
12. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: AAAI. pp. 755–760 (2004).
13. Bross, J., Ehrig, H.: Automatic construction of domain and aspect specific sentiment lexicons for customer review mining. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 1077–1086. ACM (2013).
14. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. J. Artif. Intell. Res. 50, 723–762 (2014).
15. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Comput. Intell. 22, 110–125 (2006).
16. Thet, T.T., Na, J.-C., Khoo, C.S., Shakthikumar, S.: Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In: Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. pp. 81–84. ACM (2009).
17. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th international conference on World Wide Web. pp. 519–528. ACM (2003).
18. Cui, H., Mittal, V., Datar, M.: Comparative experiments on sentiment classification for online product reviews. In: AAAI. pp. 1265–1270 (2006).
19. Tsytsarau, M., Palpanas, T.: Survey on mining subjective data on the web. Data Min. Knowl. Discov. 24, 478–514 (2012).
20. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 79–86. Association for Computational Linguistics (2002).
21. Yi, J., Nasukawa, T., Bunescu, R., Niblack, W.: Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. pp. 427–434. IEEE (2003).
22. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 115–124. Association for Computational Linguistics (2005).
23. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 513–520 (2011).
24. Moraes, R., Valiati, J.F., Neto, W.P.G.: Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Syst. Appl. 40, 621–633 (2013).

ISSN 2395-8618

25. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC. pp. 417–422. Citeseer (2006).
26. Zhu, J., Zhu, M., Wang, H., Tsou, B.K.: Aspect-based sentence segmentation for sentiment summarization. In: Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. pp. 65–72. ACM (2009).
27. Ortigosa, A., Martín, J.M., Carro, R.M.: Sentiment analysis in Facebook and its application to e-learning. Comput. Hum. Behav. 31, 527–541 (2014).
28. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. J. Am. Soc. Inf. Sci. Technol. 63, 163–173 (2012).
29. Mohammad, S.M., Kiritchenko, S., Zhu, X.: NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. ArXiv Prepr. ArXiv13086242. (2013).
30. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics. 33, 159–174 (1977).
31. Boia, M., Faltings, B., Musat, C.-C., Pu, P.: A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets. In: Social Computing (SocialCom), 2013 International Conference on. pp. 345–350. IEEE (2013).
32. Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., Kaymak, U.: Exploiting emoticons in sentiment analysis. Presented at the (2013).
33. Zhao, J., Dong, L., Wu, J., Xu, K.: Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1528–1531. ACM (2012).
34. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. 5, 1–167 (2012).
35. Aisopos, F., Papadakis, G., Tserpes, K., Varvarigou, T.: Content vs. context for sentiment analysis: a comparative analysis over microblogs. In: Proceedings of the 23rd ACM conference on Hypertext and social media. pp. 187–196. ACM (2012).
36. Adler, M., Goldberg, Y., Gabay, D., Elhadad, M.: Unsupervised Lexicon-Based Resolution of Unknown Words for Full Morphological Analysis. In: ACL. pp. 728–736 (2008).
37. Raaijmakers, S., Kraaij, W.: A shallow approach to subjectivity classification. In: ICWSM (2008).
38. HaCohen-Kerner, Y., Badash, H.: Positive and Negative Sentiment Words in a Blog Corpus Written in Hebrew. In: the 20th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Procedia Computer Science, 733-743 (2016).
39. Schutze, H., Pedersen, J.O.: A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval. Inf. Process. Manag. 33, 307–18 (1997).
40. Smadja, F., McKeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: A statistical approach. Comput. Linguist. 22, 1–38 (1996).
41. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005).

ISSN 2395-8618

A Multi-Entity Page Rank Algorithm

Chandramouli Shama Sastry, Darshan S Jagaluru and Kavi Mahesh

***Abstract*—We propose a generic multi-entity page rank algorithm for ranking a set of related entities of more than one type. The algorithm takes into account not only the mutual endorsements among entities of the same type but also the influences of other types of entities on the ranks of all entities involved. A key idea of our algorithm is the separation of prime and non-prime entities to structure the iterative evolution of the ranks and matrices involved. We illustrate the working of the proposed algorithm in the domain of concurrently ranking research papers, their authors and the affiliated universities.**

Index Terms— Multi Entity Page Rank, Mathematical Model, Evolving Stochastic Matrix, Prime Entity

I. Introduction

In this paper, we propose a mathematical model which can be used for ranking multiple interacting entities. It is widely accepted that page rank algorithm gives a meaningful and practical ranking order among a network of mutually related entities. However, the original page rank algorithm is designed for homogeneous entities and more often than not one finds it useful to rank sets of interacting or dependent entities of more than one kind in a system. The concept of ranking multiple entities at once is not entirely new and has been explored earlier, especially in ranking authors, papers and journals in a single system. However, the modifications to the page rank algorithm suggested in previous work were application specific and, as such are not readily suited to other applications. We present a generic algorithm which can be adapted to any specific application domain. We also illustrate the working of the algorithm with suitable examples and show mathematically how our algorithm is different from previous modifications. The main intuition behind the mathematical model is that tighter coupling between various entities in the ranking algorithm gives us better ranking orders. Our model ensures this by having an evolving stochastic matrix that changes iteratively along with the rank vectors of the entities.

Let's begin with an example to illustrate the intuition and motivation behind the algorithm for ranking multiple entities concurrently. Consider the problem of ranking universities purely from an academic perspective based on the research output of the universities. One would consider multiple factors for ranking: professors who work there, the research work they publish, the citations they obtain, and so on. We now face the related problem of ranking professors across universities by perhaps considering quite the same factors like the university where they work, their publications, citations, and so on. What we see here are a set of related entities whose ranks depend on each other: universities, professors, publications and perhaps others such as journals, conferences and publishers. In order to model this, we'll need a multi-entity ranking algorithm.

The outline of this paper is as follows: a brief review of previous work; a generic mathematical model for multi-entity page ranking algorithm; an example of how we can use the model for ranking authors and papers along with results in brief; a mathematical exposition of the internals of the algorithm.

II. Previous Work

Page Rank algorithm, developed by Sergey Brin and Larry Page was designed to rank web pages [1]. It is based on the random surfer model which allows a surfer to jump to a random page without necessarily following the "out-links" of the current web page. It is designed such that web pages with higher numbers of in-links or higher quality in-links, or both, are assigned higher ranks. This is a recursive algorithm that can be realized by a series of matrix multiplications. We present the core ideas and notations of the Page Rank algorithm briefly for the reader's convenience:

Let N be the number of pages to be ranked. We have a row normalized matrix H representing the directed graph of N nodes (i.e., pages) where the edges denote the hyperlinks between the pages (or any other such relation among the entities). The stochastic matrix G is then constructed as:

$$G = d \times H + \frac{1-d}{N} \times E \qquad (1)$$

where d is the damping factor describing the probability of jumps from one node to another and E is a matrix of all 1's. This matrix G should be interpreted as the matrix showing how a user may navigate from one page to another on the web. A user can either jump to a page following the links on that page or go directly to a random page. d indicates the probability that he

Acknowledgement: This work is supported in part by the World Bank and Government of India research grant under the TEQIP programme (subcomponent 1.2.1) to the Centre for Knowledge Analytics and Ontological Engineering (KAnOE), http://kanoe.org at PES University, Bangalore, India. The authors are grateful to Dr. I. K. Ravichandra Rao.

Chandramouli S Sastry and Darshan S Jagaluru are Undergraduate Students at PES University, Bangalore. Dr. Kavi Mahesh is the Dean of Research, Director of KAnOE and Professor of Computer Science at PES University.

ISSN 2395-8618

will continue following the links mentioned on the page. Empirically, d is usually set to a value of 0.85. Page ranks are computed by the formula:

$$R = R_0 \times G^n \quad (2)$$

where R is the vector containing scores (i.e., ranks) of the nodes, R_0 is the vector describing initial assignment of ranks and n is the number of iterations until convergence. R_0 is generally constructed by assuming that all nodes have equal scores. Finally, after a sufficient number of iterations *n*, nodes which have in-links from other high scoring nodes or have a lot of in-links or both, get higher scores.

Modifications suggested in previous work to adopt the page rank algorithm for multiple interacting entities have focused mainly on the ranking of authors and papers in a network [2,3,4,5,6]. We consider two main papers out of those. Zhou, Orshanskiy, Zha and Giles [6] have suggested an algorithm based on Page Rank that considers three networks – social network connecting authors, citation network connecting publications and the authorship network connecting the authors with the papers. In their algorithm, the ranks of the authors and papers are first independently computed and then coupled using intra-class or inter-class walks. That is, in terms of the original Page Rank algorithm, if the random surfer was at an author node then he could randomly jump to another author node or to a paper node. This captures the interdependency between the final ranks of authors and papers to some extent. Yan, Ding and Sugimoto [5] have also taken a similar approach where they rank journals, authors and papers together. In this work, however, the ranks of the papers, authors and journals are computed simultaneously. They create a stochastic network of papers and then create inter-entity walks between papers and journals and papers and authors. Interestingly, the stochastic matrix is dynamically updated as the ranks of authors or journals change. The formula used to update the matrix is:

$$\bar{\bar{M}} = d\bar{M} + (1-d)ve^T \quad (3)$$

where $\bar{\bar{M}}$ is the stochastic matrix, $\bar{M}$ is the adjacency matrix and e is a vector of 1s. The rationale followed is that, users don't navigate towards all papers equally; rather they jump towards papers published by reputed authors or journals or both. The vector v captures the impact of the score of the journal and the author as a metric for the probabilities of jumping. In fact, the probabilities of jumping towards a paper changes as the ranks of authors or journals change and this is reflected in the stochastic matrix. That is, higher the author score or journal score, greater the probability of a random surfer jumping towards that paper. However, this may lead to a false ranking order as average papers written by good authors get a higher score irrespective of the citations obtained by them. This is because the score assigned depends on the contents of the stochastic matrix.

We can further illustrate this drawback as follows: consider a paper that is written by very good authors and published in a reputed journal but has not yet received any cites. When we do a stochastic matrix update, its cell contents get updated and logically, all papers start giving credit to this paper (which could sometimes be even greater than the credit that those papers are giving to actually cited papers). Further, the whole process being recursive, the paper score boosts the author and journal scores which further boost the citing strength. As such, papers which may not deserve high scores end up getting them. Nevertheless, changing the matrix dynamically has its own advantages, primarily because we cannot decouple the ranking of any one of the entities from the others given their interdependencies. In our model, we evolve the stochastic matrix by considering who is citing (which, in fact is the main idea of Page Rank) rather than who is getting cited and the resulting probabilities of jumping to any node are more likely to reflect the reality of the application domain.

III. Test Data

We use a publication and citation data set that we extracted from Google Scholar to illustrate and test our algorithm. We chose all papers belonging to the subject of "Web Semantics" published between 2013-2014. The details of the dataset are as shown in Table 1.

Table 1. Data about Papers in Web Semantics in 2013-14.

Number of unique Authors	1,801
Number of Papers	1,124
Number of Citation edges (excluding self-cites)	8,294
Number of Citation edges (including self-cites)	10,192

IV. Mathematical Model

A. *Notation:*

- E=Set of all entity classes considered for ranking.
- N_i = Number of instances of entity type i.
- T_{ij} denotes the j^{th} instance of entity type i.
- O_{ij}, the inter-entity relations, where i and j are instances of different entity classes and the order of the matrix is $N_i \times N_j$. For example, i could be paper and j could be organization. These matrices need not be Boolean; they could represent real numbers as well. If we considered professors and universities, a professor can be related to multiple universities in terms of: where he studied, where he works full-time, where he works as visiting professor, and so on and we could quantify these using non-negative real valued numbers.
- L_i, the intra-entity relations, where i is the type of entity. L_i is a square matrix of order $N_i \times N_i$. These could have different semantics depending upon type of entity.

ISSN 2395-8618

- R_i – rank vector of order $1 \times N_i$ denoting scores of all instances of the i^{th} entity.

B. Prime entity

The first step is to choose an entity type P from the set E and designate it as the prime entity. The remaining entities are referred to as non-prime entities. NP_i refers to the i^{th} non-prime entity. The organization and interaction between prime and non-prime entities are as shown in Figure 1.

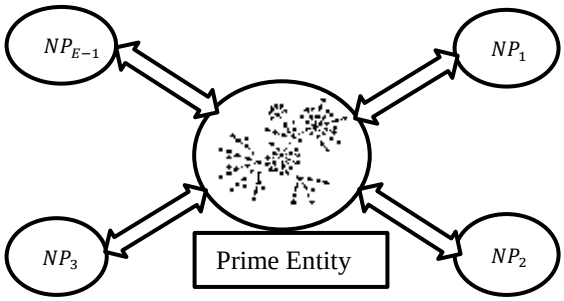

Fig. 1. Prime and Non-Prime Entities.

Prime entities are linked to one another by directed or undirected edges in a graph, whereas the non-prime entities are connected only to the prime entities. There are no edges among the non-prime entities. Prime entities serve to link the various non-prime entities. Also, the ranks of non-prime entities are influenced by the ranks of the prime entities and vice-versa.

For example, in the university ranking problem, we could consider the prime entity to be published papers. In the resulting model, papers will have a directed link between them which could denote, for example, the citation relationship. If we chose university to be prime entity, on the other hand, we could either consider an undirected graph denoting collaboration or consider a directed edge between two universities to denote that a professor who obtained his PhD from one of those universities works for the other. Choosing the prime entity determines the semantics of ranking in the chosen domain.

C. Representing the graphs

The graph connecting instances of prime entity is represented by a square matrix of order $N_P \times N_P$; this matrix is referred to as H in the future discussions. H is got by row-normalizing the matrix L_P. We then obtain the stochastic matrix G from H using (1) as described in original page rank algorithm.

In order to represent graphs linking the prime entities with the non-prime entities, we introduce:

For all $i \in E - \{P\}$,

Define $M_i = H \times O_{Pi}$ (4)

The matrix M_i quantifies the incoming links to the instances of non-prime entity i from the instances of prime entity P. Though both M_i and O_{Pi} denote and quantify the relationship between prime entity P and non-prime entity i, O_{Pi} denotes the direct relationship and M_i denotes the aggregate relationship.

In our example of university ranking, we could consider the "belongs to" relation for the inter-entity matrices between university and paper and between professor and paper. These two matrices on being multiplied with paper citation matrix would give us two matrices, each representing and quantifying the citation relationship between papers and professors and between papers and organizations thereby attributing credit to the entities.

D. Intra and Inter Entity walk

Using the matrices defined above, we can define the score of the non-prime entities in terms of the score of the prime entity as:

$$\forall\, i \in E - \{P\}, R_i = R_P \times M_i \quad (5)$$

We can also define the scores of the prime entity using the notion of page-rank (ignoring iteration numbers) as,

$$R_P = R_P \times G \quad (6)$$

R_P is initialized according to the original page rank algorithm as $R_P = \{1/N_P\ , for\ all\ T_{Pi}\ \}$.

Thus, we observe that the prime entities participate in the intra-class stochastic random walk and each of the matrices M_i help in inter-entity walk. It may be noted that this step is application independent. Continuing with our example, we can see that the model enables us to define the ranks of professors and organizations using those of papers that have cited them.

E. Building Recurrences:

This is the last and most important step of the mathematical model. Here, we modify the score vector of the prime entity based on the scores of the non-prime entities and the prime entities themselves. We use the notion of ownership and collaboration when making this modification. Note that the scores due to incoming links are already accommodated by the equations of the previous step. The general method for the modification is:

$$R_P = \alpha_0 R_P + \sum_{\forall\, i \in NP} \alpha_i R_i I_i + \beta R_P X \quad (7)$$

Where, I_i =Influence matrix$(N_i \times N_P)$ determining influence of the score of the i^{th} non-prime entity on prime entity and $X =$ collaboration matrix for determining collaboration score of a

ISSN 2395-8618

given instance of prime entity using scores of other instances of prime entity using the notion of "collaboration" and

$$\sum_{i=0} \alpha_i + \beta = 1 \qquad (8)$$

In this step, the influence matrix for each of the non-prime entities should be defined. Consider a non-prime entity W. The influence matrix I_W should define the influence that the ranks of instances of type W have on each of the instances of the prime entity. For example, assume that there are k instances of type W which influence the ranks of a certain instance T_{Pi} of the prime entity. Then, the i^{th} column of I_W should indicate the share of each of these k instances in influencing the score of T_{Pi} (other entries in the column being 0). These matrices are column stochastic. In our example, the ranks of the papers are influenced by the scores of the participating authors and organizations. Better the scores of professors and organizations, better the scores of the papers.

Collaboration matrix is used to bring in the notion of collaboration between one or more non-prime entity types with respect to a given prime entity instance. As described above, the scores of other instances of prime entity are used for factoring this in. However, in order to introduce this notion, the shares of each of the instances of the prime entity whose scores are to be considered should be based on non-prime entities. The description of this matrix is similar to that of the influence matrix with the exception that it is a square matrix since the influencing and influenced instances are of the same type. In our example, we could introduce this notion by considering the relative success of other papers when certain combinations of organizations and professors work on them. That is, any work done by a certain collaboration of organizations and professors could be as good as another. For example, works which are brought out by the collaboration of Google and Stanford involving some group of researchers from academia and industry may be at least as good as each other. In our adaptation of this model to the problem of ranking papers and authors, we show a method of including the impact of collaboration.

The factor β can be made zero if it doesn't make sense to include the collaboration factor between the non-prime entities. We have included it in the model so that certain applications could benefit by the use of this. However, if any of the αs are made zero, then the corresponding entities are effectively decoupled from the whole system. That is, the prime entity ranks are not affected by their ranks and we can compute the ranks of those instances separately. This defeats the whole purpose of multi-entity page ranking. Thus, we should not make any of these α's zero. Note that α_0 defines what percentage of the prime entity score is defined through in-links and the other α's define what percentage is defined by each of the other entities. Hence, higher the α_0, better the ranking order as a large percentage of the scores is defined through endorsements. These matrices can be static or dynamic. We give an example of one static and one dynamic in the example application that we describe later.

F. Putting it all together: pseudo-code

The scores of all the prime entities are initialized as $1/n_p$ as described in the model. The first step of the repeat-until loop involves computing the scores of the paper ranks using intra-entity stochastic walk among the network of prime entities. Following this, we compute the ranks of all the non-prime entities in terms of the prime entity. Having done this, we perform the modification step where we modify R_P to factor in the impacts of the scores of non-prime entities on the prime entity. We then normalize R_p in order to prevent arithmetic overflow. The convergence of the algorithm is then computed which determines the termination condition. The following algorithm assumes that the matrices and parameters are set as described in the preceding section.

Input:
- $E -$ Set of entities
- $P -$ Prime Entity
- $NP -$ Set of non-prime entities
- Matrix G – The stochastic matrix
- Matrices $M_i -$ Matrices for inter-entity walk
- Matrix $X -$ Collaboration matrix
- Matrix $I_i -$ Influence Matrices
- $\alpha_0, \alpha_i s \; and \; \beta$

Output:
- Ranking order of each of the entities

Procedure MERank:
Begin
  $R_P = [1/n_P \;\; ,1/n_P \, , \ldots ,1/n_P \,]$
  $\epsilon \;\; = 10^{-15}$
  Repeat:
    For every i ϵ NP
      $\mathrm{R_i} = \mathrm{R_P} \times \mathrm{M_i}$
    End
    $\mathrm{R_P^{prev}} = \mathrm{R_P}$
    $\mathrm{R_P} = \mathrm{R_P} \times \mathrm{G}$
    $\mathrm{R_P} = \alpha_0 \times \mathrm{R_P} + \beta \times \mathrm{R_P} \times \mathrm{X}$
    For every i ϵ NP
      $\mathrm{R_P} = \mathrm{R_P} + \alpha_\mathrm{i} \times \mathrm{R_i}$
    End
    Normalize R_P
    convergence = calc_convergence($\mathrm{R_P^{prev}}$,$\mathrm{R_P}$)
  Until convergence $<$ ϵ
  Return $\{R_i \; \forall \; i \; \epsilon E\}$
End

We can re-write (7) considering iteration coefficients as:

$$R_P^j = \alpha_0 R_P^j + \sum_{\forall \; i \; \epsilon \; NP} \alpha_i R_i^j I_i + \beta R_P^j X \qquad (7a)$$

In terms of R_P, we can re-write (7a) using (5) as

ISSN 2395-8618

$$R_P^j = \alpha_0 R_P^j + R_P^{j-1} \sum_{\forall\, i \,\epsilon\, NP} \alpha_i M_i I_i + \beta R_P^j X \quad (7b)$$

Observe that we use R_P^{j-1} while computing influences and R_P^j while computing collaboration impacts. The rationale is explained in the mathematical exposition.

V. Example Application Showing Ranking of Authors and Papers

A. Choosing Prime Entity

Prime Entity = Papers; Non-Prime Entities = Authors. Let entity 1 refer to Authors and entity 2 refer to Papers.

B. Representing the graphs

Intra-entity walk.

We use the paper citation graph for this. The adjacency matrix L_2 is defined as:

$$L_2[i,j] = \begin{cases} = 1.0 \text{ if paper } i \text{ cites paper } j \\ = 0.2 \text{ if papers } i \text{ and } j \text{ have at least 1 common author (self-citation)} \\ = 0 \text{ if paper } i \text{ does not cite paper } j \end{cases}$$

We can use domain-specific metrics to get better results. For example, here, we have considered the strength of a self-citation to be lower than that of a normal cite. We get the matrix H by row normalizing L_2. We constructed stochastic matrix G using damping factor value of 0.85.

Inter-entity walk

We define the matrix M_1 as the product of H and O_{21}, where O_{21} captures the ownership relation between authors and papers (order $= N_2 \times N_1$). The ownership matrix can be a real-valued matrix as well. In this example, we discriminated between the ownership influence of first author and later authors using the following idea: If author a is the k^{th} author of paper b, then $O_{21}[b,a] = \frac{2\times(n-k+1)}{n(n+1)}$, where n is the total number of authors of paper b [8].

C. Building Recurrences

Intra and inter entity walk step is common to all applications and we omit their details here.

Influence Matrix I_1

We used the transpose of the ownership matrix O_{21} as the influence matrix I_1. For this application, we chose to make it static for all iterations.

Collaboration Matrix X

For quantifying the score of collaboration of non-prime entities (authors), we developed the following formulation. For ease of explanation, consider the papers which have at least 1 common author as *co-papers* of each other. For any given paper, we partition the set of co-papers into 3 sets based on how many common authors are present [8]: A_1 the set of papers having exactly 1 common author, A_2 papers having exactly 2 common authors and A_3 papers having 3 or more common authors. For a paper B:

$$A_1 = \{\text{B and co-papers of B with 1 common author}\}$$
$$A_2 = \{\text{B and co-papers of B with 2 common authors}\}$$
$$A_3 = \{\text{B and co-papers of B with 3 or more common authors}\}$$

$\text{Collab}_{\text{Score}}[B] = s_1 \text{ x } \max(A_1) + s_2 \text{ x } \max(A_2) + s_3 \text{ x } \max(A_3)$, where $\max(A_i)$ = highest score of all papers belonging to set A_i.

The weights s_1, s_2 and s_3 are chosen such that $s_1 < s_2 < s_3$ *and* $s_1 + s_2 + s_3 = 1.0$. This ensures that the major part of the $\text{Collab}_{\text{Score}}$ is determined by the best paper in the set where most authors of the paper have collaborated. If, in case the paper B is itself the best paper that they have produced (in all the three sets), then the $\text{Collab}_{\text{Score}}$ will be same as the paper's own score, thereby ensuring that the value of such papers is not diminished, i.e., the minimum score of the $\text{Collab}_{\text{Score}}$ is same as the score of the paper. Table 2 shows values of s_1, s_2 and s_3 for different numbers of authors.

Table 2. Values of s1, s2 and s3

Number of authors	s_1	s_2	s_3
1 author	1.0	0.0	0.0
2 authors	0.25	0.75	0.0
3 or more authors	0.0625	0.1875	0.75

These computations can be represented in matrix form as:
$X = C_1 \times S_1 + C_2 \times S_2 + C_3 \times S_3$, where

$$C_x[i,j] = \begin{cases} = 1 \text{ if paper } i \text{ is the best paper of paper } j \\ \text{in the set } Ax (x = 1, 2 \text{ or } 3) \\ = 0 \text{ otherwise} \end{cases}$$

$$S_x[i,j] = \begin{cases} 0, \text{ if } i \neq j \\ s_x \text{ from Table 4 based on \# of auth of paper } i. \end{cases}$$

Note that this matrix changes in every iteration as the scores of the papers change. The changes in the entries are proportional to the convergence. The influence matrix I_1 was static whereas this matrix is dynamic.

D. Weights

As mentioned earlier, it is preferable to have α_0 greater than α_is and β to ensure that in-links dominate over ownership. In this application, we can say we prefer citations over authorship, i.e. higher score of papers shouldn't be attributed to quality of authors but rather to the quality of citations received by the papers. Secondly, we chose a higher value for β than α_1 as we

ISSN 2395-8618

wanted to give a higher weight to the collaboration factor than ownership. The values that we chose are:

$$\alpha_0 = 0.7\,, \alpha_1 = 0.1,\ \beta = 0.2$$

Results and qualitative analysis are provided in section VII.

VI. Algorithm Internals: Mathematical Exposition

The following exposition gives us an insight into the internals of the algorithm showing the evolution of the matrix. The equation for modification as explained in (7b) considering $(j+1)^{th}$ iteration is:

$$R_P^{j+1} = \alpha_0 R_P^{j+1} + R_P^j \sum_{\forall\, i \,\epsilon\, NP} \alpha_i M_i I_i + \beta R_P^{j+1} X$$

From (4), we can rewrite the above as (considering iteration numbers as well):

$$R_P^{j+1} = \alpha_0 R_P^{j+1} + R_P^j H \sum_{\forall\, i \,\epsilon\, NP} \alpha_i \mathrm{O_{Pi}} I_i + \beta R_P^{j+1} X$$

Using (6), we can substitute R_p^{j+1} on RHS as $R_P^{j+1} = R_P^j \times G$:

$$R_P^{j+1} = R_P^j \times \left(\alpha_0 G + \mathrm{H} \sum_{\forall\, i \,\epsilon\, NP} \alpha_i O_{Pi} I_i + \beta G X \right)$$

Note that if we had used R_P^{j+1} instead of R_P^j for computing influences, we'd have had a H^2 term associated with the second term; this causes R_i to be ahead by one time-step. Hence, we use R_P^j for computation of R_is. We can simplify this using (1) as:

$$R_P^{j+1} = R_P^j \times \left(d \times \left(\alpha_0 H + H \sum_{\forall\, i \,\epsilon\, NP} \alpha_i O_{Pi} I_i + \beta H X \right) + \frac{1-d}{N} \alpha_0 E + \frac{1-d}{N} \beta E X \right) \quad (9)$$

We can rewrite this as:

$$R_P^{j+1} = R_P^j \times \left(dH' + \frac{1-d}{N} (\alpha_0 + \beta) E \right) \quad (10)$$

by replacing $\alpha_0 H + H \sum_{\forall\, i \,\epsilon\, NP} \alpha_i O_{Pi} I_i + \beta H X$ with H' and simplifying $E \times X$ as E because matrix X is column-stochastic i.e., sum of all columns of X is 1. Thus, we can now see how the stochastic matrix evolves as the ranks change. We also see that the probabilities of jumping to any of the nodes is equal$\left(= \frac{1-d}{N} (\alpha_0 + \beta) \right)$. The matrix H' of this iteration $(j+1)$ will be the H of next iteration $(j+2)$. Also, by substituting (9) in (5), we can see that the non-prime entities get their ranks based on all the other entities.

VII. Results

We executed our algorithm on different configurations of the parameters α_0, $\alpha_1$and β. We describe four important configurations here. The results in the form of author ranks and paper ranks are shown in Tables 3 and 4 respectively for the four cases:

Case 1: The configuration followed in this case is α_0=1, α_1=0 and β=0. This configuration is same as computing the scores of papers using normal page rank algorithm and then computing the scores of the authors using these. We can see that the ranks of authors and papers are linearly related.

Case 2: The configuration followed in this case is α_0=0.7, α_1=0.3 and β=0. This configuration considers the scores of the owning authors along with the citations a paper has got for computing the ranks. We observe that top six authors remain in the same positions. However, authors like LK, NH and MH have got a higher rank than previous configuration. The reason is they've been cited by many papers authored by ACNN. However, in case 1, he didn't receive much credit because these citations were considered as self-citations. But, in this case, even though the citations were considered self-citations, each of the citing paper's score itself was enhanced by ACNN's and his co-authors' scores, which caused them to move higher up in the ranking order. Also, paper B which is cited by paper A, got 1 rank lower than paper C, which is cited by many more papers having ACNN as author.

Case 3: The configuration followed in this case is α_0=0.7, α_1=0 and β=0.3. This configuration considers the collaboration scores of every paper along with the paper's own score. Here, there's no direct coupling between author and paper scores. However, the collaboration matrix brings in the coupling between the two entities. At the very first look, we find that the top ranking author is now IHW instead of ACNN. IHW's paper D has been cited by ACNN and friends who have maintained a consistent top record in all their papers. Hence, he's got a higher score. We find that this ranking order is little too strict given the strict nature of the technique used for computing the collaboration scores.

Case 4: The configuration followed in this case is α_0=0.7, α_1=0.1 and β=0.2. In this case, we combine the good features of case 2 and case 3. We chose to give a little higher share to the collaboration scores of the paper than the owning author scores as we get a more meaningful ranking order using this.

We analyzed the results qualitatively and saw how our algorithm could factor in collaboration and dependence between inter-entity scores into ranking. For a more complete evaluation, we computed h-index - a popular metric used for the

ISSN 2395-8618

purposes of ranking in the problem domain chosen - of all the authors within our data set and compared it with the author ranking our algorithm produces. We generated the score-distribution graphs for both page rank and h-index for comparison. While we plotted the h-index scores as they were, we scaled the page rank scores 100 times and considered top 80 percentile scores for ease of visualization. Figure 2 shows the distributions of h-index and page-rank when applied on our data set.

Table 3. Top 24 Author Ranks for the Four Cases

	Author	Case 1 Rank	Case 2 Rank	Case 3 Rank	Case 4 Rank		Author	Case 1 Rank	Case 2 Rank	Case 3 Rank	Case 4 Rank
1	ACNN	1	1	7	1	18	MAM	18		20	
2	MV	2	2	13	2	19	CG	19		21	
3	SA	3	3	8	6	20	MS	20	22	18	22
4	JL	4	4	9	7	21	RN	21	16	11	17
5	AZ	5	5	10	8	22	DMH	22	24	3	9
6	KL	6	6	12	5	23	RB	23		4	10
7	IHW	7	10	1	3	24	HH	24		5	11
8	DM	8	11	2	4	25	CU		13		23
9	LK	9	7		13	26	PC		14		24
10	NH	10	8		14	27	VL		15		
11	MH	11	9		15	28	DG		17		
12	VC	12	12		16	29	H		21		21
13	BH	13	18	14	18	30	AM		23		
14	WR	14	19	15	19	31	PM			6	12
15	JN	15	20	16	20	32	GDG			22	
16	MH2	16		17		33	OM			23	
17	JDF	17		19		34	DC			24	

Table 4. Top 10 Ranks of Papers in the Four Cases

Paper	Authors	Case 1 Rank	Case 2 Rank	Case 3 Rank	Case 4 Rank
A	SA, JL, AZ, ACNN	1	1	3	1
B	MV	2	3		5
C	ACNN, KL	3	2	4	3
D	DM, IHW	4	6	2	4
E	ACNN	5	4		6
F	ACNN	6	5		7
G	ACNN, LK, NH, MH	7	8		8
H	ACNN, KL, VC	8	9		9
I	JDF, MAMP, CG	9		7	
J	DC, GDG, DL, ML	10		5	
K	DG, ACNN		7		
L	RB, HH, DMH, PM		10	1	2
M	BH, WR, JN, MH2			6	10
N	MS			8	
O	PC, DS, SP, TC			9	
P	OM			10	

ISSN 2395-8618

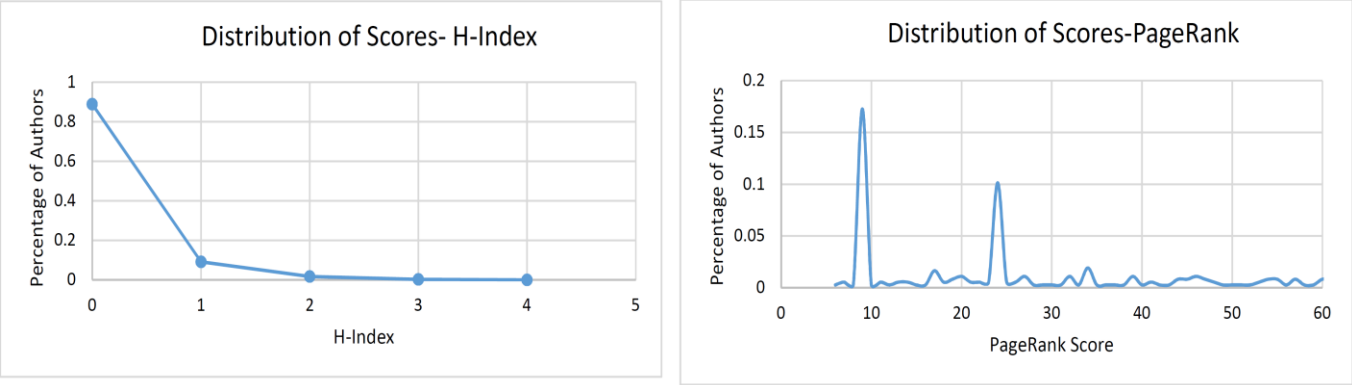

Fig. 2. Distributions of h-index and multi-entity page-rank scores

A clear observation is that h-index maps all 1,801 authors to just 5 distinct h-index scores, whereas page-rank assigns a larger distribution of scores to the authors. This can be attributed to the fact that the proposed model can account for factors like collaboration and inter-entity dependence and give a finer score allocation and method or ranking. We also report a 75% increase in the number of distinct scores when multi entity page rank algorithm is used in comparison to the original page rank (got by using the configuration defined in Case 1). From the h-index distribution, we can infer that the data set under consideration has relatively newer authors and their papers. Thus, the model which we have proposed considers multiple factors and is capable of producing a finer and practical ranking order even in cases where h-index fails to provide a clear distinction among the authors.

The results show that our multi-entity page-rank algorithm, while accounting for a number of factors as against conventional page rank algorithm, can produce practical and meaningful ranking orders of multiple related entities.

VIII. Discussion

As we had earlier mentioned, the semantics of the ranking is defined by the way we choose the prime entity and the relations between them. For example, in the university ranking problem, if we choose the prime entity to be papers and consider the citation graph, the results are purely in the academic perspective. That is, works which are well-cited have a positive influence on the university ranks and once, universities which have produced more novel works get a higher rank. The same logic holds for professors as well. However, if we considered a directed graph of universities, where the links indicate that professors who've obtained PhDs from one university serve for the other, universities professors from good universities or both, will get a higher rank. This graph can be inverted to mean that universities producing lot of good professors get a good rank. As another example, we could also consider the undirected graph denoting collaboration between the professors (or universities) as the prime entity graph. Yan and Ding in their work [7] have reported good results by applying page rank on an undirected graph of authors denoting co-authorship**.**

Given the variety of choices that one could make, the success of the algorithm depends on how the influence matrices and collaboration matrix are defined in consideration of the available data and the semantics of the domain.

Conclusion

In this paper, we have proposed a generic mathematical model of a multi-entity ranking algorithm which employs basic concepts of page rank, whose parameters can be set appropriately depending on the application. We've also provided a mathematical derivation showing the internals of the algorithm which is important in designing a successful ranking model. We have given examples throughout the paper illustrating the use of various parameters, which aid in designing a model which is well-suited to the application at hand.

References

[1] Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1), 107-117.

[2] Laure Soulier, Lamjed Ben Jabeur, Lynda Tamine and Wahiba Bahsoun (2012). On Ranking Relevant Entities in Heterogeneous Networks Using a Language-Based Model. *Journal of the American Society for Information Science and Technology*. Volume 64, Issue 3, pages 500-515.

[3] Ming Zhang, Sheng Feng, Jian Tang, Bolanle Ojokoh and Guojun Liu (2011). Co-Ranking Multiple Entities in a Heterogeneous Network: Integrating Temporal Factor and Users' Bookmarks. *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation: 13th International Conference on Asia-Pacific Digital Libraries*, (ICADL 2011), pages 202-211.

ISSN 2395-8618

[4] Tehmina Amjad, Ying Ding, Ali Daud, Jian Xu, Vincent Malic(2015). Topic-based Heterogeneous Rank. *Scientometrics*. Volume 104, Issue 1, pages 313-334.

[5] Yan, Erija, Ding, Ying and Sugimoto, Cassidy R. (2011). P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*. Volume 62, Issue 3, pages 467–477.

[6] Zhou, Ding, Orshanskiy, Sergey, Zha, Hongyuan and Giles, C. Lee (2007). Co-Ranking Authors and Documents in a Heterogeneous Network. *IEEE International Conference on Data Mining* (ICDM 2007), pages 739-744.

[7] Yan, Erija, and Ding, Ying (2011). Discovering author impact: A PageRank perspective. *Information processing & management,* 47(1), 125-134.

[8] Chandramouli Shama Sastry, Darshan S. Jagaluru, and Kavi Mahesh (2016) Author ranking in multi-author collaborative networks. *COLLNET Journal of Scientometrics and Information Management*, 10(1), 21-40.

ISSN 2395-8618

Application of Adaptive PD Control to a Biomechanics System Known as Exoskeleton

Salvador Juarez-Lopez [1] and Ezequiel Molinar-Solis[1] and Adrian Juarez-Lopez[2] and Marco A. Villeda-Esquivel[1]

1 Universidad Autónoma del Estado de México, México.
2 Instituto Tecnológico de Orizaba, México.

***Abstract*—The control in biomechanical systems has become an active field of biomedical engineering. This paper describes a design of adaptive control to achieve desired in the system defined as exoskeleton moves, the adaptive method is located is selected taking into account energy savings because the adaptive controller is based on a scheme of variable gain in time Proportional and Derivative respect to PD control. The structure of adaptive gain is determined using a type of control Lyapunov function. The adaptation law uses velocity estimation based on a robust exact differentiator (RED) implemented as a variation of Super- Twisting algorithm. The derivative adaptive proportional controller is evaluated on simulated exoskeleton structure. The set of simulations considers the presence of an external disturbance. The controller proves efficient to counter the effects of external mechanical system. Proposed controller performance was superior to standard proportional derivative controller, and has been shown in this study.**

***Index Terms*—Biomechanical system; Adaptive Control PD; Exoskeleton.**

I. Introduction

THE development of exoskeletons has boomed in the last 20 years and because it has helped humans to solve a myriad of problems, especially in the biomedical area. In particular exoskeletons have been an interesting line of research and widely studied by many researchers in the military, industrial and medical area, the latter being the greatest social impact, because it provides a direct benefit to patients with mobility problems neuromuscular. Here, to potentiate the processes of physical rehabilitation of persons with motor disabilities are developing new robotic devices such as exoskeletons [8][9].

Exoskeletons are a kinematic chain which engages externally individuals whose joints and links correspond to the joints of the human body which tries to emulate. The main feature of these mechanisms is direct contact between the user and the exoskeleton, which transfers the mechanical power through information signals[1].

If your paper is intended for a conference, please contact your conference editor concerning acceptable word processor formats for your particular conference.

In the area of biomedical, exoskeletons can help mainly in the field of rehabilitation, and this is achieved by controlling the movements of the patient through a process of stimulation between mechatronic systems and patient. The design and control of an exoskeleton is focused on achieving patient develop a uniform to finally get a steady gait rehabilitation[10]. This is classified as static and dynamic walk, wherein the first center of gravity is maintained in a second region and its center of gravity is not kept in the same region. Both are capable of balancing a person, obtaining stability. This paper presents the design of an exoskeleton and development of control algorithms based on a control known as adaptive control [3].

II. Notation

The following notation was used in this study: $\mathbb{R}^n$ represents the vector space with n-components, τ is used to define the transpose operation, $\|k\|$ is used to define the Euclidean norm of is $k\epsilon\mathbb{R}^n$. $\|k\|_H^2 \coloneqq k^T Hk$ is the weighted norm of the real valued vector $k \in \mathbb{R}^n$ with weight matrix $H > 0, H = H^T,\ H \in \mathbb{R}^{n\times n}$. The matrix norm labeled as $\|D\|^2, D \in \mathbb{R}^{n\times n}$ is defined as the maximum eigenvalue of the matrix D. If two matrices $N \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{n\times n}$, fulfills $M > N$, that means that M-N is a positive definite matrix. The symbol $\mathbb{R}^+$ represents the positive real scalars. The symbols $I_{n\times n}$ and $0_{n\times n}$ were used to represent the identity matrix $I\in \mathbb{R}^{n\times n}$ and the matrix formed with zeroes of dimension $n \times n$. This is just some harmless text under a subsection.

III. State space formulation of Exoskeleton

Whereas the mechanical structure of an exoskeleton, this can be formulated in an equation of state space,

$$M(q)\ddot{q} + C(q,\dot{q}) + g(q) + \Delta(t,q,\dot{q}) = u(t)$$

Where $q \in \mathbb{R}^n$ is the vector of generalized coordinates, M is the inertia matrix, C is the matrix of Coriolis and centrifugal forces, g is the gravitational force term Δ is the term of uncertainty y u denotes the vector of controllable forces provided by the torque required to move the actuators. The control input u is assumed that some functions is given by known feedback. Note M (q) is invertible, where $M(q) = M(q)^T$ and is strictly positive definite.

ISSN 2395-8618

Using the state variable representation of the mechanical structure (1), the second order nominal model presented above can be represented as follows:

$$\frac{d}{dt}x_a(t) = x_b(t)$$

$$\frac{d}{dt}x_b(t) = f\big(x(t)\big) + g\big(x(t)\big)u(t) + \Delta(x(t),t) \quad (1)$$

The vector x_a represents the position in each degree of freedom of the exoskeleton; the associate vector x_b is the corresponding velocity. Finally, the function Δ represents the uncertainties and perturbations. In this paper, it is assumed that

$$\|\Delta(x(t),t)\|^2 \le \eta_0 + \eta_1\|x\|^2, \eta_0, \eta_1 \in \mathbb{R}^+$$

The control structure was proposed following the adaptive PD scheme. This class of control model obeyed

$$u(t) = \Big(g\big(x(t)\big)^{-1}\Big)\Big(k_p(t)\Big)e(t) + k_d(t)\frac{d}{dt}e(t)$$

Where

$$x = \left[e^\intercal \quad \frac{de}{dt}\right]^\intercal, e = [x]^\intercal$$

$$\frac{de}{dt} = [x_b]^\intercal$$

The mechanical nature of exoskeleton is used here to consider that a nonlinear system described by a feasible distributed second order nonlinear differential equation can be used for representing it mathematically.

He drif term $f: \mathbb{R}^{2n} \to \mathbb{R}^n$ is a Lipschitz function. The following assumption is considered valid in this study.

Assumption 1. The nonlinear system (1) is controllable.
Based o the previous fact, the input associated term $g: \mathbb{R}^n \to \mathbb{R}^{n\times n}$satisfies.

$$0 < g^- \le \|g(k)\|_F \le g^+ < \infty, k \in \mathbb{R}^n \quad (2)$$

It is evident that matrix g(z(t)) is invertible $t \ge 0$.
Assumption 2. The nonlinear function f $(\cdot)$ is unknown but satisfies the Lipschitz condition

$$\|f(x) - f(x')\| \le L_1\|x - x'\| \quad (3)$$

I the previous inequality, $\le x, x' \in \mathbb{R}^{2n}$and $L_1 \in \mathbb{R}^+$.

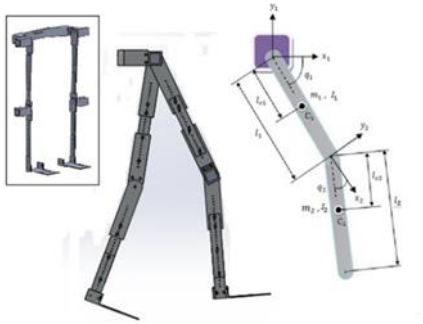

Fig.1 Represents the exoskeleton and its free body diagram to obtain its mathematical model of the system by the Euler-Lagrange method.

The Euler-Lagrange generate a mathematical model of the system is represented by the following two equations:

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + g(q) + \Delta = u$$

$$\begin{bmatrix} M_{11}(q) & M_{12}(q) \\ M_{21}(q) & M_{22}(q) \end{bmatrix}\ddot{q} + \begin{bmatrix} C_{11}(q,\dot{q}) & C_{12}(q,\dot{q}) \\ C_{21}(q,\dot{q}) & C_{22}(q,\dot{q}) \end{bmatrix}\dot{q} + \begin{bmatrix} g_1(q) \\ g_2(q) \end{bmatrix} + \Delta(t,q,\dot{q}) = u$$

IV. Controller Structure

A PD controller is designed using the assumption regarding $e(t)$ and $\frac{dy}{dx}(t)$ are measured simultaneously where e is the tracking or the regulation error. This is not the regular case in real building mechanical structure represented in Figure 1. Otherwise, an important resources investment. Therefore, in classical in classical literature, one can find two important solutions: to construct an observer or using a …first order filter to approximate the error derivative. The first one requires the system structure (that is in this paper is assumed to be unknown because the presence of external perturbations and internal uncertainties) and in the second case, the derivative approximation is usual poor, especially if the output information is contaminated with noises. One additional option is considering a class of RED that can provide a suitable and accurate approximation of the error derivative. Super Twisting Algorithm (STA) has demonstrated to be one of the best RED in several times[4][5].

IV.1 Super-Twisting Algorithm

In counterpart of some others second order sliding modes algorithms, the STA can be used with systems having relative degree one with respect to the chosen output Levant (1993). The STA application as a RED is described as follows. If $w_1 = r(t)$ where $r(t) \in \mathbb{R}$ is the signal to be differentiated, $w_2 = \frac{dr}{dt}(t)$ represents its derivative and under the assumption of $\frac{dr}{dt}(t) \le r^+$, the following auxiliary equation is gotten $\frac{dw_1}{dt}(t) = w_2(t)$ and $\frac{dw_2}{dt}(t) = \frac{d^2r}{dt}(t)$. The previous set of differential equation is a state representation of the signal r(t).

The STA algorithm to obtain the derivative of r(t) looks like

$$\begin{aligned} \frac{d}{dt}\overline{w_1}(t) &= \overline{w_2}(t) - \lambda_1|\hat{w}_1(t)|^{\frac{1}{2}}\text{sign}\big(\hat{w}_1(t)\big) \\ \frac{d}{dt}\overline{w_2}(t) &= -\lambda_2\text{sign}\big(\hat{w}_1(t)\big) \\ \hat{w}_1 &= \overline{w_1} - w_1 \\ d(t) &= \frac{d}{d(t)}w_1(t) \end{aligned} \quad (4)$$

where $\lambda_1, \lambda_2 > 0$ are the STA gains. Here $d(t)$is the output of the differentiator Levant[2].

$$\text{sign}(z) := \begin{Bmatrix} 1 & \text{if} & z > 0 \\ [-1,1] & \text{if} & z = 0 \\ -1 & \text{if} & z < 0 \end{Bmatrix} \quad (5)$$

ISSN 2395-8618

IV.2 PD controller with the Super-Twisting Algorithm

A single adaptive PD controller is applied over each section of the building-like mechanical structure. This is a class of ATMD. Each adaptive PD controller proposed in this study obeys the following structure

$$u_i(t) = -k_{1,i}(t)e_i(t) - k_{2,i}(t)d_i(t) \qquad (6)$$

where e_i is x^a. The gains in the PD controller are determined by

$$k_1(t) = g_i^{-1}\big(x_a(t)\big)\big(k_1(t) + k_1^*\big)$$

(7)

$$k_2(t) = g_i^{-1}\big(x_a(t)\big)\big(k_2(t) + k_2^*\big)$$

With $k_1(t)$ and $k_2(t)$ are time varying scalars adjusted by a special tracking error dependent adaptive law described by the following ordinary differential equations:

$$\frac{d}{dt}\bar{k}_{1,i}(t) = -\pi_1^{-1}e_i(t)M_a^T P_2 E_i(t)$$

(8)

$$\frac{d}{dt}\bar{k}_{2,i}(t) = -\pi_2^{-1}e_{i+n}(t)M_b^T P_2 E_i(t)$$

Where $\pi_{1,i}$ and $\pi_{2,i}$ are free parameters to adjust the velocity of convergence for the adjustable gains. In (7), the parameters $k_{1,i}^*$ and $k_{2,i}^*$ are positive constants. The matrices Ma and Mb are given by $M_a = [1\ 0]^T$ and $M_b = [0\ 1]^T$: Additionally, the term $E_i = [e_i\ \ e_{i+n}]$. The matrix $P_{2,i}$ is positive definite and it is presented in the main statement of the main theorem of this article.

The variable $d_i(t)$ is obtained from the following particular application of the STA as RED:

$$\frac{d}{dt}\bar{x}_a^i(t) = \bar{x}_a^i(t) - \lambda_1|\hat{x}_1(t)|^{\frac{1}{2}}\mathrm{sign}\big(\hat{x}_1(t)\big)$$
$$\frac{d}{dt}\bar{x}_b^i(t) = -\lambda_2\mathrm{sign}\big(\hat{x}_a(t)\big) \qquad (9)$$

$$\hat{x}_a^i(t) = x_a^i - \bar{x}_a^i$$

Considering that displacements on building-like structures are small and considering the assumption 1 and 2, it is easy to get that $\left\|\frac{d}{dt}x_b^i(t)\right\| \le h^*$ where h^* is a finite positive scalar.

The following extended system describes the complete dynamics of the error signal in close-loop with an adequate implementation of (4) and the controller proposed in (6):

$$\frac{d}{dt}x_a^i(t) = x_b^i(t)$$
$$\frac{d}{dt}x_b^i(t) = f\big(x^i(t)\big) - \bar{k}_{1,i}(t)e_i(t) - \bar{k}_{2,i}(t)d_i(t) + \Delta^i(x^i(t), x^{i+1}(t), x^{i-1}(t), t)$$

$$\frac{d}{dt}\hat{x}_a^i(t) = \hat{x}_a^i(t) - \lambda_{1,i}|\hat{x}_a(t)|^{\frac{1}{2}}\mathrm{sign}\big(\hat{x}_a(t)\big)$$

$$\frac{d}{dt}\hat{x}_b^i(t) = -\lambda_{2,i}\left(\hat{x}_a(t)\right) - \frac{d}{dt}x_b^i(t)$$

(10)

$$\frac{d}{dt}\bar{k}_{1,i}(t) = -\pi_{1,i}^{-1}e_i(t)M_a^T P_{2,i}E_i(t)$$

$$\frac{d}{dt}\bar{k}_{2,i}(t) = -\pi_{2,i}^{-1}e_{i+n}(t)M_b^T P_{2,i}E_i(t)$$

The following section shows the main result of this paper. The theorem introduced in that section gives a constructive way to adjust the gains of the STA and it provides the applicability of using the adaptive gains for the PD controller.

IV.3 Convergence of the adaptive PD controller

The stability of the $e = 0$ is justified by the result presented in the following theorem:

Theorem 1. Consider the nonlinear system given in (1), supplied with the control law (6) adjusted with the gains given in (7) and the derivative of the error signal obtained by means of equation (9), if there exist a positive scalar i and if the gains are selected as $\lambda_{1,i} > 0$, $\lambda_{2,i} > 0$the next Lyapunov inequalities always have a positive definite solution $P_{1,i}$.

$$A_{1,i}^T P_{1,i} + P_{1,i}A_{1,i} \le -Q_{1,i}$$

$$A_{1,i} = \begin{bmatrix} -\lambda_{1,i} & 1 \\ -\lambda_{2,i} & 0 \end{bmatrix}; Q_{1,i} = Q_{1,i}^T > 0; Q_{1,i} \in \mathbb{R}^{2\times 2} \qquad (11)$$

then for every positive value of L1 satisfying equation (3) and positive value of h^+, there exist positive gains $\bar{k}_{1,i}, \bar{k}_{2,i}$ such that if the Riccati equations given by

$$P_{2,i}\big(A_{2,i} + \alpha_i I\big) + \big(A_{2,i} + \alpha_i I\big)P_{2,i}^T + P_{2,i}R_{2,i}P_{2,i} + Q_{1,i} \le 0 \qquad (12)$$

have positive definite solution $P_{2,i}$ with

$$A_{1,i} = \begin{bmatrix} 0 & 1 \\ -k_{1,i}^* & -k_{2,i}^* \end{bmatrix}; R_{2,i} = \Lambda_{a,i} + \Lambda_{b,i}$$

(13)

$$Q_{2,i} = 4\lambda_{\max}\{\lambda_{b,i}^{-1}\}I_{2\times 2} + \bar{\Lambda}_{a,i}; \bar{\Lambda}_{a,i} = L_1\,\Lambda_{a,i}$$

$$\Lambda_{a,i}, \Lambda_{b,i} > 0, \text{and symmetric}, \Lambda_{a,i}, \Lambda_{b,i} \in \mathbb{R}^{n\times n}, \alpha_i \in \mathbb{R}^+$$

and if the adaptive gains of the PD controller are adjusted by (8), thus the trajectories of

$$E^T = \left[x_1^a, \dots, x_n^a, x_1^b, \dots, x_n^b\right]$$

ISSN 2395-8618

$$\lim_{t\to\infty} E^{\intercal}(t)P_2E(t) \leq \sum_{i=1}^{n}\frac{\gamma_i}{\alpha_i} \qquad (14)$$

where

$$P_2 = \begin{bmatrix} P_{2,1} & 0_{2\times2} & 0_{2\times2} \\ 0_{2\times2} & P_{2,2} & 0_{2\times2} \\ 0_{2\times2} & 0_{2\times2} & P_{2,n} \end{bmatrix} \qquad (15)$$

and $\gamma_i = 2\lambda_{max}\{\Lambda_{b,i}^{-1}\}(h_i^{+} + 2\Delta_{+}^{*} + \eta_{0,i})$

V. Numerical Results

Fig.2 Represents the Exoskeleton that it was built and where the control tests were performed.

When the PD is calculated adaptive controller for the system, the derivative obtained by the STA provides some advantages. The robustness of STA is applied as a wrapper performs best for any driver applied to second order systems when the only information available is the output signal.

Then the first part of the numerical simulations is employed to evaluate the performance of the STA as RED.

The derivative of exoskeleton like positions is compared with the information provided by the measurements obtained directly from the simulation of the system presented in Figure 1 and the derivative of reference signal.

Calculations were done in Matlab software with the following parameters of Table 1.

Table 1 exoskeleton parameters were obtained with the Solidworks software

i	$m_i[Kg]$	$l_i[m]$	$l_{ci}[m]$	$I_{xxi}[Kgm^2]$	$I_{yyi}[Kgm^2]$	$I_{zzi}[Kgm^2]$
1	0.734	0.48	0.24	257.681	136.610	121.386
2	0.564	0.48	0.24	429.689	86.953	345.264

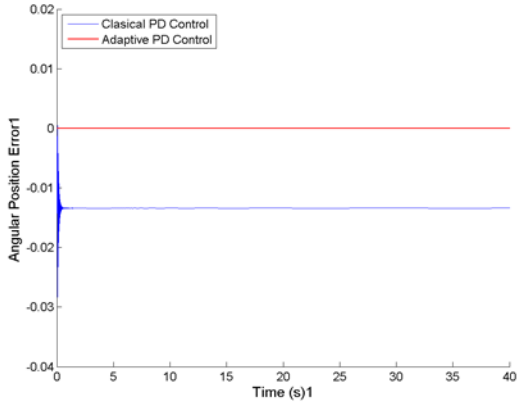

Fig.3 Evaluate the adaptive PD control and classical PD for the board of the exoskeleton located in the thigh.

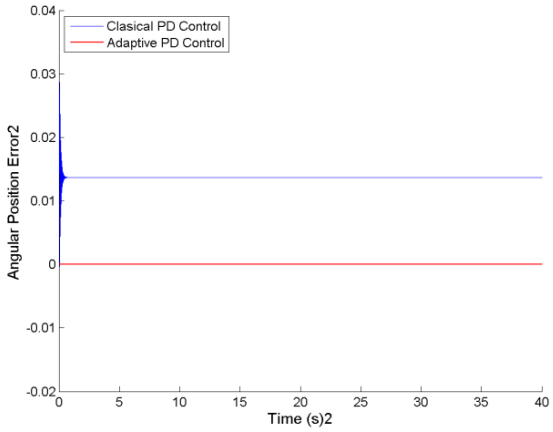

Fig.4 Evaluate the adaptive PD control and classical PD for the board of the exoskeleton located in the calf.

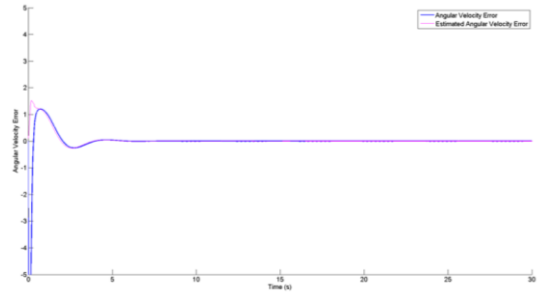

Fig.5 Evaluate the error adaptive PD control and classical PD for the board of the exoskeleton located in the thigh.

VI. Conclusion

An adaptive output based controller based on the proportional derivative controller was implemented to the exoskeleton. The controller was fed with the information of the velocity estimated by a RED based on the application of the super twisting algorithm. The closed loop controller forced the ultimate boundedness of the tracking errors to a region around the origin. A special class of Lyapunov function was the main tool for obtaining the adaptive gains of the PD controller as well as the convergence of the STA used as RED.

Simulation observed in Fig. 3 shows that the algorithm STA

ISSN 2395-8618

is zero, while the adaptive PD gets that stability in a close to zero but is never zero, so does the board located in the calf fig. 4, therefore the algorithm STA obtains better control for this system.

REFERENCES

1. Levant, A. (1993). Sliding order and sliding accuracy in sliding mode control. International Journal of Control, 58(6), 1247—1263.

2. Levant, A. (1998). Robust exact differentiation via sliding mode tecnique. Automatica, 34(3), 379--384.

3. Poznyak, A. (2008). Advanced Mathematical Tools for Automatic Control Engineers: Volume 1: Deterministic Systems. Elsevier Science.

4. Davila J., Fridman L. and Levant A. (2005) Second-Order Sliding-Mode for Mechanical System, IEEE Transactions on Automatic Control, 50(11),1785-1789.

5. J. A. Moreno and M. Osorio. (2008) A Lyapunov approach to second-order sliding mode controllers and observers, 47th IEEE Conference on Decision and Control.

6. G. Bartolini, A. Pisano, E. Punta, andE. Usai, "A survey of applications of second-order sliding mode control to mechanical systems," Int. J. Control, vol. 76, pp. 875–892, 2003.

7. H. Sira-Ramirez, "Dynamic second-order sliding mode control of the hovercraft vessel," IEEE Trans. Control Syst. Technol., vol. 10, no. 6, pp. 860–865, Nov. 2002.

8. Kazerooni, H. ; Racine, J.-L. ; Lihua Huang ; Steger, R., "On the Control of the Berkeley Lower Extremity Exoskeleton (BLEEX)", Robotics and Automation, ICRA 2005. Proceedings of the 2005 IEEE International Conference.

9. Vanderborght, B. ; Verrelst, B. ; Van Ham, R. ; Vermeulen, J. "Dynamic Control of a Bipedal Walking Robot actuated with Pneumatic Artificial Muscles", Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference.

10. Zoss, A.B.; Kazerooni, H. ; Chu, A.,"Biomechanical design of the Berkeley lower extremity exoskeleton (BLEEX)", Mechatronics, IEEE/ASME Transactions on (Volume:11 , Issue: 2), 2006.

ISSN 2395-8618

Planning Techniques to Compose Experimental Protocols in Civil Engineering

Ingrid-Durley Torres, Juan Felipe Muñoz-Fernández, Jaime A. Guzmán-Luna

***Abstract*—This paper presents a study that shows how planning techniques in Artificial Intelligence can be used in experimental protocols composition. Also, it presents how it is necessary to represent the experimental protocol models under ASTM standard with the emerging technologies as ontologies. The main goal of this paper is to discuss the way composition process can be used to meet the needs of the users. Additionally, this paper also searches to define the way the Civil Engineering domain must normalize its experimental practices, if we want the results of experimental activities to be reproducible and reusable.**

***Index Terms*— planning techniques, experimental protocols composition, standard ASTM, ontologies, and Civil Engineering**

I. Introduction

THE experimental practices of Civil Engineering are standardized in documents that are defined by ASTM International (American Association for Testing and Materials) [2]. Some of these documents describe, as an experimental protocol [8, 10, 14], how individual (atomic) tests of laboratory must be done for the testing of specific materials, which are used in processes of the Civil Engineering domain. For ex-ample, materials such as soils, metals, among others are subject to these laboratory tests.

In this domain, at the same time, there exist more complex processes as the design of pavements [1], which requires a soil study in its preliminary phase. This soil study should be composed of the individual execution of the experimental protocols de-fined in documents ASTMD0422, ASTM4318, ASTM698, just to mention some of them. This is to say, each experimental protocol, seen in an isolated way and independent from the context of the study, is not very useful for the user of the domain of Civil Engineering [7], but as a whole, they can lead to structure more complex processes that cannot be assisted by only one protocol, but for the composition of some of them. Nevertheless, the task is not trivial, since the protocols that should be obliged to participate in the process of composition must be identified accurately, according to the desired process. Similarly, it is required to know the order of execution of every protocol because, in some cases, the result of the execution of a proto-col constitutes the input to execute the next one or a subsequent one (functional requirement). It is also necessary to identify the protocols that can be executed using the resources of the laboratory (non-functional requirements). For example, the protocols defined in documents ASTM2974 and ASTMD4643 measure the percentage of humidity of a soil sample. However, the first one describes the test using a drying oven, while the second one describes the test using a microwaves oven.

In this way, a process of civil engineering [12] as a soil study can be characterized in a protocol composed of several atomic protocols, which have been obtained as a subordinate sequence of elements described from their functional and non-functional requirements.

The composition is an emerging solution, which is successfully applied in other domains such as Web Services [11] and learning routes [12]. It corresponds to a process, which is attained on one of its most promising lines by implementing planning techniques, and the automatic interaction of pre-existing elements to generate new elements that suit the specific needs of a user, when those needs cannot be satisfied by the preexisting elements.

In this paper, the concept of composition of experimental protocols used in the domain of Civil Engineering is presented, using planning techniques in artificial intelligence. In this composition, to consider the functional and non-functional aspects can be useful to overcome the adaptation and contextualization problems of the users aim, formulated as a complex process, and becoming a significant contribution for this domain. As an emergent contribution, it is proposed to model the semantic representation of the documents ASTM as well as the hierarchy of the main composed processes of the civil engineering area. Nevertheless, as a study case for this work, the representation will be focused on presenting the

Manuscript received October 9, 2001. (Write the date on which you submitted your paper for review.) This work was supported in part by the U.S. Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials.

I-D T. Author, is MsC. and researcher professor at Institución Univeristaria Salazar y Herrera, currently a PhD student at the Universidad Nacional de Colombia, headquarter Medellin (corresponding author to provide phone: 57-4-4600700; Ext: 2218; e-mail: idtorresp@unal.edu.co).

J. F. M-F. Author, is Esp. and MsC. student author, at the Universidad Nacional de Colombia, headquarter Medellin (e-mail: jfmunozf@@unal.edu.co).

J. G-L. Author, is PhD. and MsC A, is professor associate at the Universidad Nacional de Colombia, headquarter Medellin; leader of the research group SINTELWEB (e-mail: jaguzman@unal.edu.co).

ISSN 2395-8618

preliminary phase of the design of pavements.

This paper is organized as follows: the second section describes the relative concept of ASTM standards and the Civil Engineering process. The following section shows the representative model used for describing the experimental protocols under ASTM standard, with ontologies. In the fourth section, we describe the composition process, including the functional and non-functional requirements of the users, and the conclusions are discussed in the fifth section.

II. Civil Engineering Domain

One of the key aspects for the users involved in the domain of Civil Engineering: engineers, laboratory workers, auditors, technicians, etc., is the knowledge that they must have when applying the ASTM standards. Such knowledge will enable them to read and understand the documents of a project, as well as prepare them; and perhaps most importantly, this knowledge will enable them to avoid mistakes in quality assurance and the control of the construction and manufacturing [7]; The ASTM standards can be classified in five big groups: testing methods, material specifications, recommended practices, nomenclature and guides. For the scientific interest, the first group of this work is focused on standards of testing methods, which deter-mine the properties and characteristics of the material; this is the reason why they correspond to experimental practices carried out in laboratories [13]. At present, these standards are applied, hand documented, and written in natural language; and only in some cases, they are partially transcribed to a digital file. It is the applied ASTM standard which ultimately makes it an experimental protocol, that is to say "a sequence of tasks and operations that are represented as recipes providing the description of the processes step by step, described in natural language. Such sequence, in experimental research, is considered a fundamental unit of knowledge" [3]. In daily life, the laboratory workers document every protocol with valuable information, which is related to the used resources, the data employed, the implemented activities, the obtained results, the generated conclusions, together with some other related annotations. However, if a user wants to accompany a process, as in the case of pavement design. According to the profile, this user must analyze a specific aspect of the result of a sequence of protocols that represent the phase of the addressed study (as shown in Fig. 1). This task must be done manually, aiming to read and understand the documentation (written) involved in the project, and related in this phase. So, in the case of the engineer, he can concentrate on the results of the tests, while an auditor can, using the same information, identify the calculations and tests (just to mention some of them). Even if another general user wants to implement only those protocols that are used by certain equipment, this user must use the manual search. As quoted, each user has different specific interests, even using the same documentation. This is the reason why the normalization of the documented information is valuable to strengthen processes such as the indexation, search, recovery, and their composition, in order to ease the human labor.

III. Representation Model

The formal description of the experiments is necessary in order to obtain an efficient analysis of the investigation, and to allow the exchange of scientific results, turning into a fundamental part of the practice of science. Thus, representing the domain of the experimental protocols, of the field of civil engineering semantically, allows modeling the scientific investigation of the area. Such representation must be unambiguous, reach an agreement, and it must contain all the common and needed elements to reuse the experiments; no matters for whom or for what are they required.

A. Experimental Protocol Ontology

For the goal of proposing a semantic representation that supports the needs of the experimental protocols in civil engineering, it is proposed to implement an ontology.

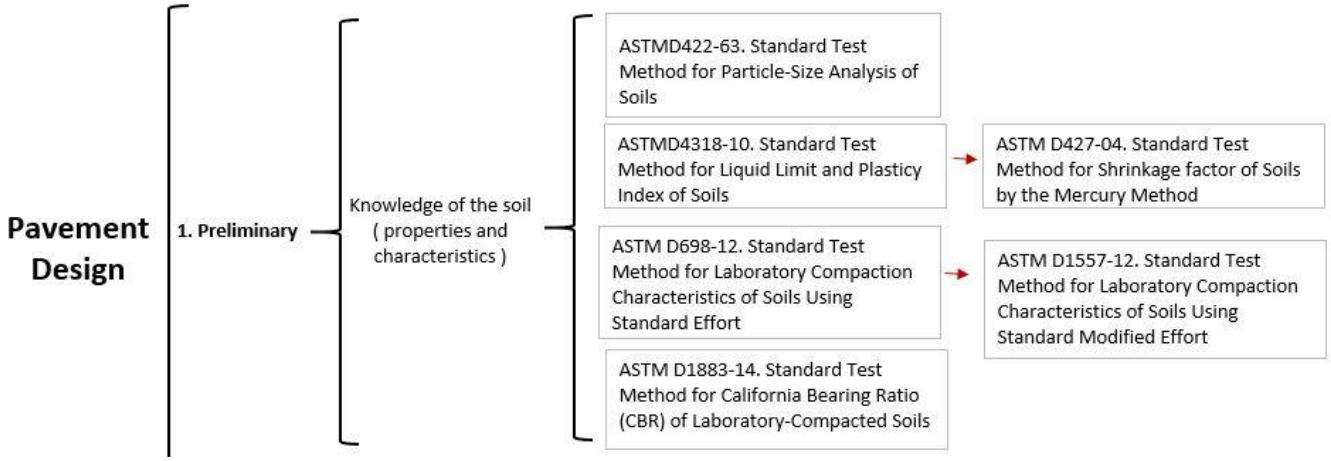

Fig. 1 Portion of the Process of Pavement Design (Preliminary Phase)

ISSN 2395-8618

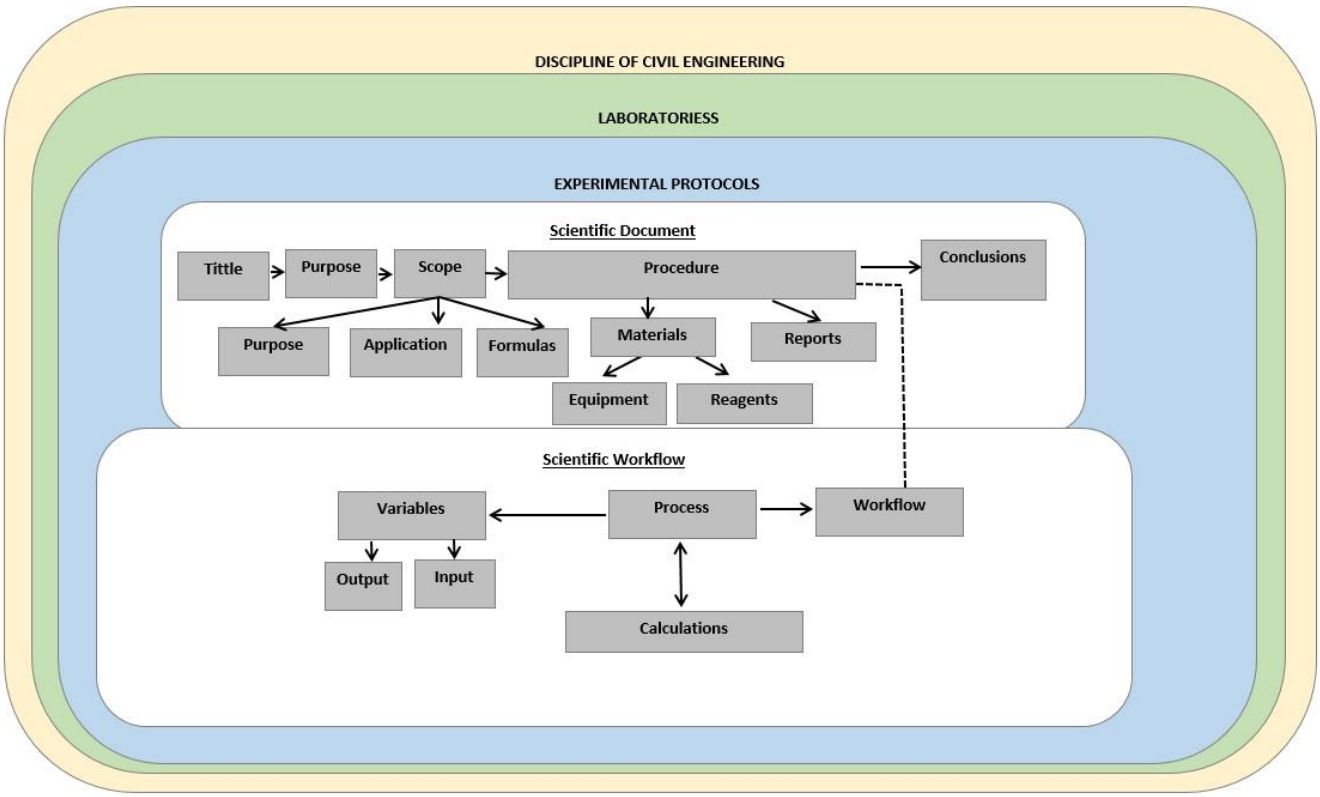

Fig. 2. Experimental protocol, as document and as scientific workflow, inspired [8]

In this stage, it is possible to notice that the experimental protocol is composed of two key elements: a documental structure that represents the steps of the scientific meth-od¸ and a description modeled through a flow of procedures (known as scientific workflow) [4, 5]. The previous description reflects the SMART Protocol [8], which is designed for the domain of the biotechnology of the vegetable plants. The goal is to model the experimental protocol in two parts (see Fig. 2), separating the static information model (called Document Scientific in Fig. 2) from the dynamic model (called Scientific Workflow in Fig 2, too).

The first being represented as a scientific document that registers the basic textual information of the methodology of investigation (modeled as ontology after), and with which it is possible to model non-functional requirements (tittle, objective, purpose, results and other metadata of documenting the experiment, from the steps of the scientific method). Whereas, the second represents the protocol as a flow of processes, commonly known as workflow, i. e. the set of step in an experiment. A Workflow can be defined as a directed graph in which the nodes are implemented by human ("mixed the sample") or software programs (calculations or simulations). This element is used for representing the functional requirements (the Fig. 4 show an example).

The intention of using the ontology was the clearest task so far: The classification and reuse of experimental knowledge generated in the laboratories of the Civil Engineering field. Thus, the most focused recommendation was about reusing the existing ontologies, since by the time of considering the reuse of standards of ontological type in this field, proposals under the scheme of modeling experiments were done. In order to enumerate the most important terms of the domain, the decision aims to reuse the SMART Protocol ontology, obviously extended and adapted to the needs of the experimental civil engineering; thereby generating a PECE-Protocol Experimental Civil Engineering Ontology as document (see Fig. 3). PECE show a model ontology, which enumerates of the set of requirements to be considered, when an experimental proto-col in civil engineering is represented. In this case there are elements that are continuant and others are currents. Between the first are all entities that exists in full at any time in which it exists at all, persists through time, under this specification are all elements of scientific document (see Fig. 2). But in this case these elements are sections of the *class:documentpart;* in the second share, the occurrences entities, are all ele-ments that exist in a given moment of time, scientific workflow (see Fig. 2) *class:workflow*.

For space limits the Fig. 3 shows only a portion of subclasses *class:Entity* and *Class:Continuant* besides of the some elements *subclassOf:DocumentPart* development in Protégé software.

A. Complex Processes Domain Ontology

Retaking the idea of dealing with the semantic heterogeneity among users (scientist, engineer, laboratory workers, and auditor), the protocol composition model goes beyond simple ontological domain modeling, and it includes semantic formalization of a complex process.

ISSN 2395-8618

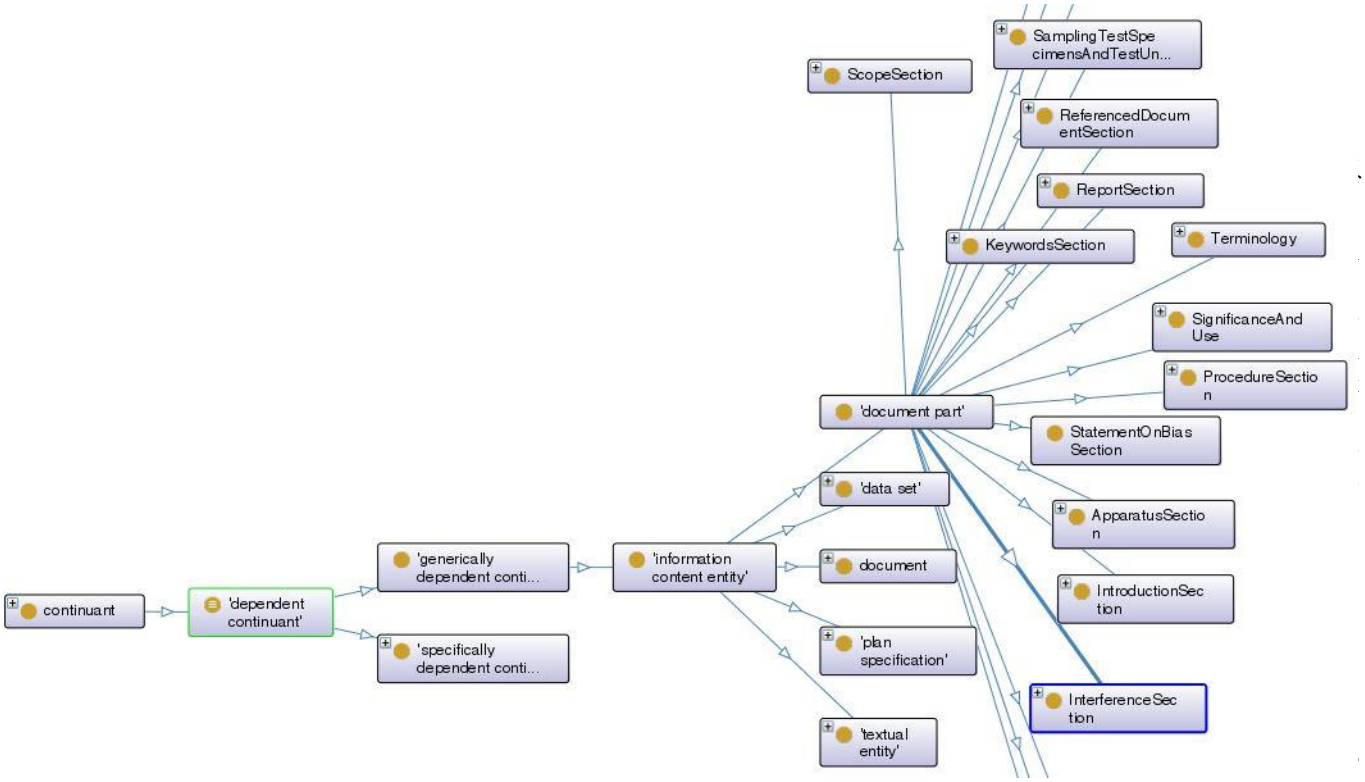

Fig. 3. Portion of Ontology PECE as Document

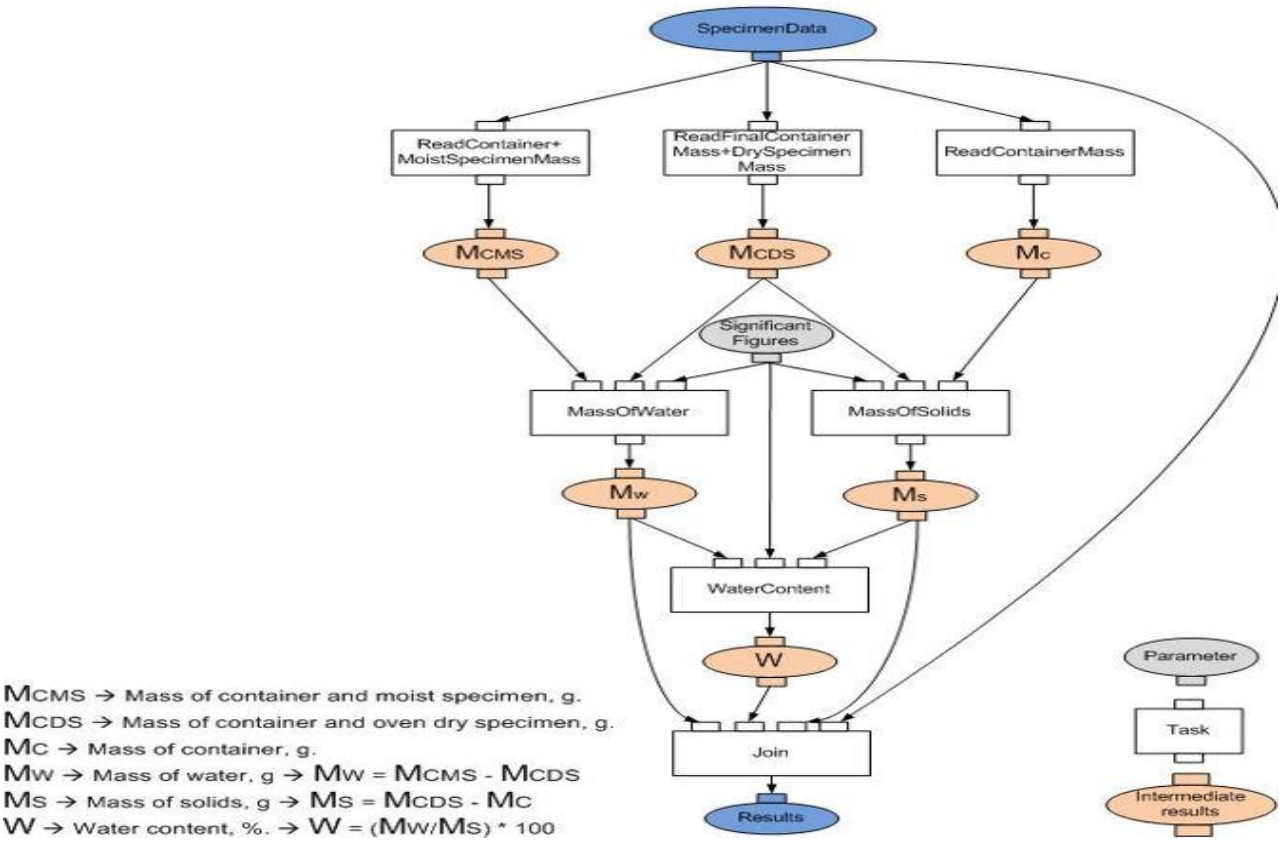

Fig. 4. ASTMD2216-5122 Protocol as Workflow

ISSN 2395-8618

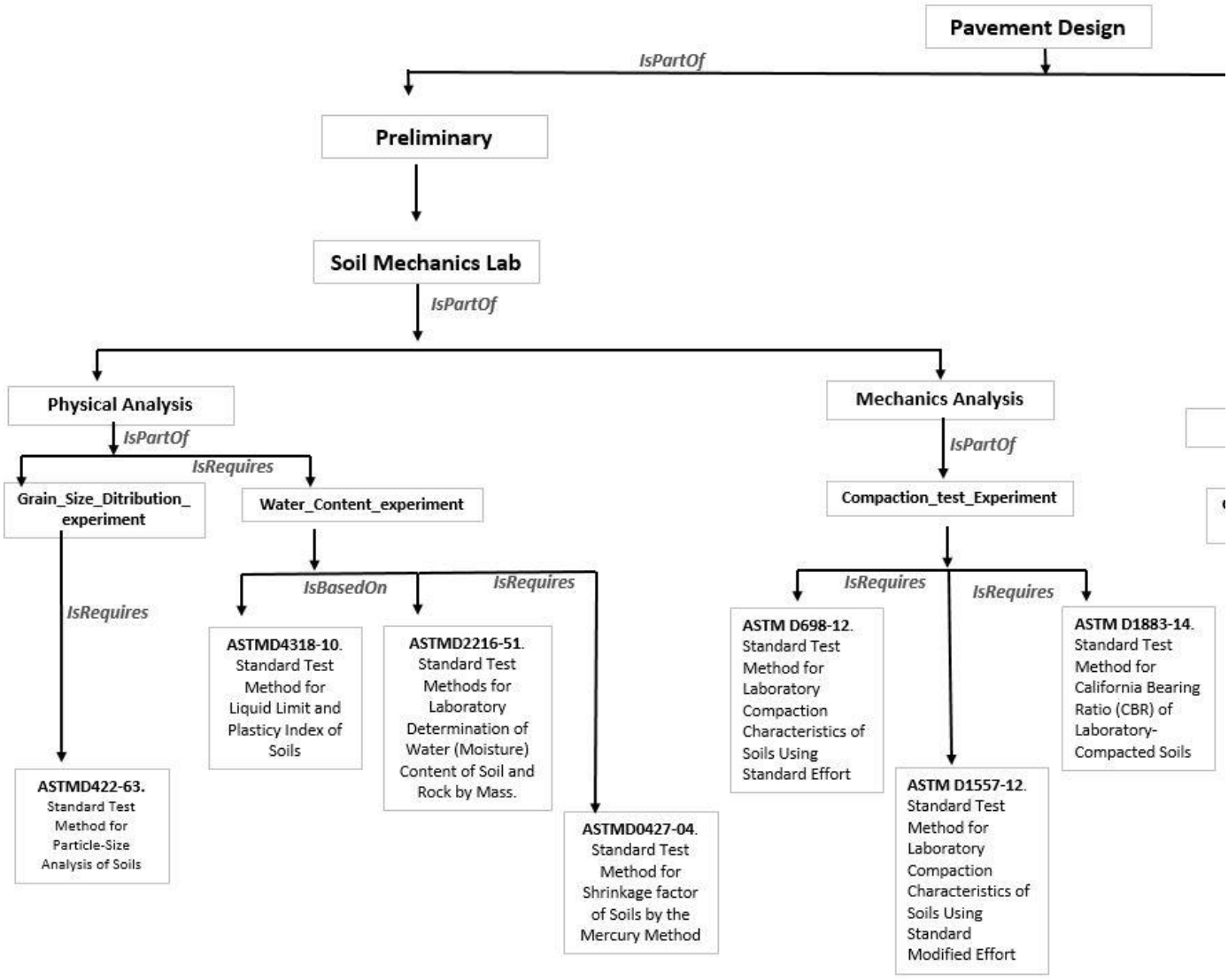

Fig. 5. Portion of Model Complex Process Domain Ontology

It defines a concept like *OWL:Concept* [15] which belongs to an *OWL:Class*. Consequently, it is possible to build a network of concepts depending on the cause-effect relation, which may take place among the various phases of the complex process (See Fig. 5). The idea is to model each complex process as a hierarchical network of tasks where the branches correspond to the last level phases mentioning experimental protocols to be achieved with the application of each ASTM. Order these values enable to define a linear order among the actions, or among the internal or external processes for the satisfaction of the objectives of the user. i) *IsBasedOn*, provides order relations among protocols (ASTM Instanced) and the processes. It can provides order relations among protocols (ASTM Instanced) and the processes; it can be defined as disjunctive "O", since a process can need to know the outcomes provided by a "Y" protocol, also obtained by a "Z" protocol in order to achieve the knowledge goals of the "Y or Z" specific process. This is, then, at least one of the "*OR"* type must have been completed in order to access the following protocol. ii) *Requires*, reports dependencies among type; this happens when a process needs the outputs from another process. This is a hard requirement, that is to say, it must be obtained before addressing the next one. Hierarchy: iii) *IspartOf:* this concept describes a hierarchical structure among the elements of the process. The aim is to model the domain of the complex processes of Civil Engineering, as shown in Fig. 1, through this ontology.

V. The Composition Process

The composition of an experimental protocol, from the point of view of a user's requirement, is directly related to the formal specification of a planning problem, which includes collections of actions with pre-requisites and results (preconditions and effects respectively). Inspired in [3, 11]. When an action supports another, this generates a causal link between them, meaning that the preceding action is finished before starting a succeeding action. With these actions, it is possible to establish a plan of the actions which will be transferred to users (Scientist, engineer, auditor, etc.) from an initial state of knowledge of a complex process in Civil Engineering, to a state in which formulated goals are achieved

ISSN 2395-8618

by users themselves. Expressed analogically, with a civil domain, users express their requirement depending on one or various concepts of domain specific knowledge; these concepts are part of a complex process (as shown in Fig.2). This requirement must be achieved from a state of a user's initial knowledge of that domain (which is even considered null, for the specific case in which an individual knows nothing of a domain); also, including at the same time considerations regarding preference of the non-functional requirements in this case expressed as metadata of ontology as document. Thus, the experimental protocol compound for this case constitutes a (plan) sequence of protocol experimental associated to knowledge domain (the specific Civil complex process, i.e. pavement design, review database buildings, beam analysis, to name a few), and customized for a user.

As an attempt to transfer both specifications, in one sole model (Planning Domain Definition Language PDDL [9]), was born, which was expressed by the (*S, So, G, A, R*) tuple, where *S* is a set of all the possible states, defined by the all the concepts of the complex process map. $So \in S$ denotes the initial state of the problem, and is related with one or more, state of the complex process map (concepts of the domain Ontology); $G \in S$ denotes the goal state, defined by the user´s specific complex process, that, he must be achieve. This process, are expressed by a set of concepts associated to the domain ontology of the Civil complex process (Fig. 5), in which the composition system will try to search. A

is a set of actions which represent the protocols which the composer must consider to change from a user state of knowledge, to another state of knowledge, and the translation relation $R \in S \times A \times S$ defines the precondition and effects to execute each action.

As it is deduced from the tuple, *S* states are defined depending on the concepts associated with the complex process model; this is why, this stage can only be achieved once that, at least one of the experimental protocols of the process modeling has been defined. This composition process, in this case is done in two phases: (i) in order to select the sequence of protocols, that allow to reach the goal (expressed as one or more concept of Domain Ontology). In this phase, the system works with the functional requirements, this is the workflow, expressed by, its inputs and/or outputs. (ii) in order to consider the best adapted assignments to a given user preferences, it is only possible with the non-functional requirements; this is the metadata of the protocol as document.

In the first case, each protocol (ASTM Instanced) is translated to a planning domain as an action, where preconditions will be defined as two types (i) some of knowledge directly associated to Complex Process Domain Ontology concepts and to the assignments; and (ii) some precedence associated to be represented by is required and *isBasisFor*. The effects themselves correspond to the statement that says that the concrete protocol is already known and that task specific has already been carried out (*task_done*). Another important aspect, apart from the order of the protocol on the concept map of the ontology of domain of complex processes, is the sequence of input and output variables associated to the workflow (seen as black box [3]) of each protocol. These are also part of the domain description, incorporating as pre-conditions and effects of knowledge. The above types represent a specification problem, already in PDDL. Fig. 6 represents a portion a domain of pavement. In this Fig., it is possible to appreciate the way, the ASTM are converted in the Domain File expressed as PDDL actions; while Fig. 7 represents the file problem, also in PDDL, in this case it includes all the objects, the initial state, and the goal user´s specification.

```
(define (domain ROPavimentos)
(:requirements :strips :typing)
(:types ro inputvar outputvar process language - object)
(:constants MCMS MCDS MC MW MT MD N PL A R G G1 V VO Mmtf Wtf Pc Mt1 Mmd V1
            K Wc Yw W1 W2 B MassAsreceivedSpecimen
            MassoftheOvenDriedSpecimen - inputvar
            W LL PI P D SL DrySoilMass WetSoilMass Mdtf Pf Pm Pd Yd Wsat
            Gs A1 MoistureContentAirDriedSample AshContent - outputvar
            en es - language
)
(:predicates
 (inputvar_known ?e - ro ?i - inputvar)
 (outputvar_known ?e - ro ?o - outputvar)
 (task_ASTM_D2216_done ?process - process)
 (task_ASTM_D4318_done ?process  - process)
 (task_ASTM_D0422_done ?process  - process)
 (task_ASTM_D0427_done ?process  - process)
 (task_ASTM_D0698_done ?process  - process)
 (task_ASTM_D136_done ?process  - process)
 (task_ASTM_D2974_done ?process  - process)
 (null ?process  - process)
)

(:action ASTM_D2974
    :parameters (?e - ro ?i - inputvar ?o - outputvar ?process - process)
    :precondition (and (inputvar_known ?e MassAsreceivedSpecimen)
                       (inputvar_known ?e MassoftheOvenDriedSpecimen) (null ?process))
    :effect (and (outputvar_known ?e W)(outputvar_known ?e MoistureContentAirDriedSample)
                 (outputvar_known ?e AshContent)(task_ASTM_D2974_done ?process))
)
```

Fig. 6. Portion of PDDL domain (pavement)

In the second case, the convenience of the experimental protocol is selected according to the preferences of the user, and it is achieved by building an abstract plan (see Fig. 7), achieved in the previous phase. After the discovery of each protocol in the repository of the protocol, it is necessary to filter each one according to the preference of each user. This is achieved by a semantic matching, each preference with the corresponding metadata. In this case, the selection preferences metadata are shown to the user for him to select the ones of his interest. For instance, the user can only require to analyze the pavement protocols of the preliminary phase, carried out by Universidad Nacional de Colombia (*Metadata Author*), in July 2016 (*Metadata Date*); it may even request the system to report only the ones, whose materials use a microwaves. As it is possible to observe, the preferences of the user, based on interests, must be paired with the metadata of the protocol as a document. For this reason, once the plan is built, as pointed in Fig. 8, the system must search only those elements whose matching is precise to the one formulated by the user. Such specification is expressed as an "*and*", this is to say, each and every preference must be mapped in the repository of protocols. In addition, this matching exercise can show results of two types: i) that one does not exist (or any protocol), or ii) there exist more than one protocol matching the preferences.

ISSN 2395-8618

```
(define (problem P1)
    (:domain ROPavimentos)
    (:objects
        ASTM - ro
        pavimentos - process
    )
    (:init
        (null pavimentos)
        (inputvar_known ASTM MCMS)
        (inputvar_known ASTM MCDS)
        (inputvar_known ASTM MC)
        (inputvar_known ASTM MW)
        (inputvar_known ASTM MT)
        (inputvar_known ASTM MD)
        (inputvar_known ASTM N)
        (inputvar_known ASTM PL)
        (inputvar_known ASTM A)
        (inputvar_known ASTM R)
        (inputvar_known ASTM G)
        (inputvar_known ASTM G1)
        (inputvar_known ASTM V)
        (inputvar_known ASTM VO)
        (inputvar_known ASTM Mmtf)
        (inputvar_known ASTM Wtf)

    )
    (:goal
        (and (outputvar_known ASTM A1)(task_ASTM_D136_done pavimentos)
```

Fig. 7. Portion of PDDL problem

In the first case, the user must change his preferences in order to do a new selection, and if necessary, to build an abstract plan again (it can be considered alternative protocols defined under the relation *IsBasedOn*, being previously ignored). If in this case, the terms that make up the original search of the user do not coincide with the existing concepts in the context base, the agent returns a fault in the request. In the second case, there exists more than one protocol matching the preferences. Conversely, after finding these metadata, the system returns to an RDF document [9] with a list of the triplets related to information about all metadata.

```
2:(astm_d0698 astm ?i ?o pavement)
          6   -> (task_astm_d0427_done pavement)
          3   -> (task_astm_d4318_done pavement)
          0   -> (inputvar_known astm yw)
          0   -> (inputvar_known astm wc)
          0   -> (inputvar_known astm k)
          0   -> (inputvar_known astm v1)
          0   -> (inputvar_known astm mmd)
          0   -> (inputvar_known astm mt1)
          0   -> (inputvar_known astm pc)
          0   -> (inputvar_known astm wtf)
          0   -> (inputvar_known astm mmtf)
;

1:(astm_d136 astm ?i ?o pavement)
          2   -> (task_astm_d0698_done pavement)
          0   -> (inputvar_known astm b)
          0   -> (inputvar_known astm w2)
          0   -> (inputvar_known astm w1)
;

bindings =
{ ?e(3) ?e(5) ?e(6) ?e(4) ?e(2) ?e(1) } == astm
{ ?process(6) ?process(5) ?process(4) ?process(3) ?process(2) ?process(1) } == pavement
;
orderings = { 2<1 3<1 3<2 3<6 4<1 4<2 4<3 4<5 4<6 5<1 5<2 5<3 5<6 6<1 6<2 }
Time: 0
```

Fig. 8. Portion of a plan built

VI. *Discussion and Conclusions*

The technology of the experimental protocols semantically modeled is recent. It emerged from the application of the ideas coming from the Semantic Web field in the experimentation area. This technology is based on adding formal and comprehensible semantic contents, by computer systems, to the description of the capacities of the protocols, in such a way that software entities can process this information the same way that a human would. Although there are very well represented models, it has never been this necessary that each domain adjust its representation to its own requirements; if reusing automatically the results of the experimentation is being sought. Simultaneously, this paper has presented an approximation about the way in which the planning technique can be incorporated to the design of a protocols com-position in the domain of Civil Engineering; specifically when modeling the ASTM standard under functional and non-functional requirements.

The need for new alternatives that enable the reuse of documented results of an experimental research, better adapted to the needs of the role of a user, has motivated the application of planning techniques of Artificial Intelligence in this field. In this regard, all the efforts have been oriented to adapt the way to represent the domain of the protocols semantically, and to identify the best way to achieve eloquence from the existing planners, allowing to deal with these characteristics.

The described model is part of a representation, composition, and execution plat-form of experimental protocols, being the reason why we are working on the best way to adapt the execution process to the functional and non-functional requirements associated to the workflows, and to the semantic representations respectively. Nevertheless, the main effort is represented in the work demanded by the sensitization of the community, when sharing the results of the researches; if it is intended that e-science starts to permeate the scientific experimentation with its benefits in the do-main of Civil Engineering.

Acknowledgment

The research reported in this paper is supported from the investigation thesis “A Integrated Model of logical reasoning, and planning techniques in Artificial Intelligence for Automatic Composition of Research Object based in Experimental Protocols in the Area of Civil Engineering” and “A recommender model of experimental protocols based on user usage context”. Both funded under the Universidad Nacional de Colombia, headquarters Medellin.

References

[1] Asphalt Institute. A Basic Asphalt Emulsion Manual. Manual Series No. 19 (MS-19), (1979).

[2] ASTM Standard. American Society for Testing and Materials - ASTM International. Available: https://www.astm.org/ (1996).

[3] Barreriro D and Albers P.: Approaches of Web Services Composition Comparison Between BPEL4WS and OWL-S. In Proceedings of ICEIS'05(International Conference on Enterprise Information Systems), May 23-27,pp. 208-213, Miami-USA, (2005).

[4] Bechhofer S., Soiland-Reyes S. and Belhajjame K. Workflow Lifecycle Management Initial Requirements. , SWAT4LS, London, (2011)

[5] Beco S., Cantalupo B., Giammarino L., Matskanis N. OWL-WS: a workflow ontology for dynamic grid service composition. First International Conference on e-Science and Grid Computing, 1-1 July (2005).

[6] Corcho O. Research Semantic S-cience. http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/researchareas/3-semanticscience/index.html.

ISSN 2395-8618

[7] Delatte N. Las Normas y la Ingeniería en las Aulas Universitarias. Magazines & Newsletters / ASTM Standardization News. Entrevista Mayo/Junio (2009)

[8] Giraldo O, Garcia A, and Corcho O.: SMART Protocols: Semantic Representations for Experimental Protocols. 4th Workshop on Linked Science - Making Sense Out of Data (LISC2014). 19th or 20th October. Riva del Garda, Trentino, Italy (2014).

[9] McDermott, D., and the AIPS-98 Planning Competition Committee (1998). PDDL-the planning domain definition language. Tech. rep., www.cs.yale.edu/homes/dvm.

[10] Soldatova L. and King R.. An ontology of scientific experiments. Journal List. J. R. Soc Interface. Vol. 3 No. 11. doi: 10.1098/rsif.2006.0134. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1885356/. (2006).

[11] Torres I-D, Guzmán J..: Reactive Planning to Compose Learning Routes in Uncertain Environments. CISSE - University of Bridgeport. December 12-14, (2013).

[12] Torres I-D, Guzmán J.: Applying Case-Based Learning to Improve the Efficiency in the Web Service Compositions. International Journal of Engineering and Technology. vol. 6, no. 3, pp. 227-233, (2014)

[13] Universidad Nacional de Colombia. Laboratorios Universidad Nacional de Colombia, heardquare Medellín. Facultad de Minas. Retrieved Abril, 2014, from http://www.minas.medellin.unal.edu.co/dirlab/index.php/laboratorios. (2000).

[14] Verdes-Montenegro L.. E-Science for the SKA, WF4Ever: Supporting Reuse and Reproducibility in Experimental Science RadioNet Advanced Radio Astronomy, Commissioning Skills and Preparation for the SKA Manchester: November 15th (2012)

[15] OWL, Web Ontology Language, https://www.w3.org/OWL/.

ISSN 2395-8618

Multi-day-ahead traffic accidents forecasting based on Singular Spectrum Analysis and Stationary Wavelet Transform combined with Linear Regression: A comparative study

Lida Barba and Nibaldo Rodríguez

***Abstract*—Numerous statistical and mathematical methods have been developed in order to explain the complexity of nonstationary time series. Singular Spectrum Analysis (SSA) and Wavelet Transform (WT) are two potent theories with different mathematical foundations that have been used in several applications with successful results; however in most studies SSA and WT have been presented separately, then there is a lack of systematic comparisons between SSA and WT in time series forecasting. Consequently the aim of this work is to evaluate the performance of two hybrid models, one is based on SSA combined with the Autoregressive model (SSA-AR), and the other is based on Stationary Wavelet Transform combined with AR (SWT-AR). The models are described in two stages, the first stage is the time series preprocessing and the second is the prediction. In the preprocessing the low frequency component is obtained, and by difference the high frequency component is computed. Whereas in the prediction stage the components are used as input of the Autoregressive model. The empirical data applied in this study corresponds to the traffic accidents domain, they were daily collected in the Chilean metropolitan region from 2000 to 2014 and are classified by relevant causes; the data analysis reveals important information for road management and a challenge for forecasters by the nonstationary characteristics. The direct strategy was implemented for 7-days-ahead prediction, high accuracy was observed in the application of both models, SWT-AR reaches the best mean accuracy, while SSA-AR reaches the highest accuracy for farthest horizons.**

I. Introduction

Singular Spectrum Analysis (SSA) and Wavelet Transform (WT) are two potent methods with different mathematical foundations that have been successfully applied in time series analysis; nevertheless, in the literature review there is a lack of systematic comparisons between SSA and WT even more in forecasting. Singular Spectrum Analysis is a nonparametric spectral estimation method which is used to decompose a time series into a sum of components such as trend, cyclical, seasonal, and noise. SSA is defined in four steps, embedding, Singular Value Decomposition (SVD), grouping and diagonal averaging, which are summarized in Decomposition and Reconstruction [1]. The beginning of the SSA method is attributable to Loève (1945)[2], Karhunen (1946), [3] and Broomhead-King (1986) [4]. Some researches have taken advantage of the SSA flexibility to apply it in diverse fields; favorable results were obtained in climatic series [5], [6], energy [7], industrial production [8], tourist arrivals [9], trade [10], among others. Although the SSA flexibility allows its usage in a wide range of relatively short time series, there is a lack of standard methods to select the window length, which is a principal parameter in the decomposition.

On the other hand the wavelet decomposition is a popular method of nonstationary time series analysis in the time-frequency domain. Successful results have been obtained in a wide number of applications such as hydrology [11, 12], biological signals [13], common energy consumption [14], financial market [15], [16], marketing [17], among others. The wavelet analysis provides spectral and temporal information in different spatial and temporal scales. The Continuous Wavelet transform (CWT) [18] and the Discrete Wavelet Transform (DWT) [19, 20] are used to obtain a representation form of a time series. CWT calculates wavelet coefficients at every possible scale, which requires a significant amount of computational resources and it generates redundant information. Whereas DWT calculates the wavelet coefficients based on discrete dyadic scales; DWT reduces the computational complexity and generally non-redundant information, however it is prone to shift sensitivity, which is an undesirable feature in forecasting [21]. Stationary Wavelet Transform (SWT) is based on a nonorthogonal multiresolution algorithm for which the DWT is exact [22], besides SWT is shift invariant. The SWT decomposition is dependent of two parameters, the wavelet function and decomposition levels, regrettably as in SSA, there is no standard methods to take decisions about these parameters configurations.

This paper contribution is a systematic comparison of hybrid models based on Singular Spectrum Analysis combined with Autoregressive model (SSA-AR) and Stationary Wavelet Transform combined with the same Autoregressive model

Manuscript received on August 7, 2016, accepted for publication on October 26, 2016, published on June 30, 2017.

Lida Barba is with the Facultad de Ingeniería, Universidad Nacional de Chimborazo, 060102, Riobamba, Ecuador, and the Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, 2362807, Valparaíso, Chile.

Nibaldo Rodríguez is with the Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, 2362807, Valparaíso, Chile.

ISSN 2395-8618

(SWT-AR) for multi-step ahead forecasting of nonstationary time series. A daily time series of injured in traffic accidents is used to evaluate the hybrid models; the data were collected in the Chilean metropolitan region (Santiago) from year 2000 to 2014 [23]. This paper is organized as follows. Section II describes the methodology used to implement SSA and SWT. Section III presents the prediction based on components. Section IV shows the efficiency criteria. The Results and Discussion are described in Section V. Finally Section VI concludes the paper.

II. Methodology

The forecasting methodology is described in two stages, preprocessing and prediction (Fig. 1). Singular Spectrum Analysis (SSA) and Stationary Wavelet Transform (SWT) are implemented in the preprocessing stage. The aim of SSA and SWT is to decompose an observed signal x in components of equal size and different dynamic, in this case of low and high frequency, c_L and c_H respectively. In the prediction stage the Autoregressive (AR) model is implemented to predict the components.

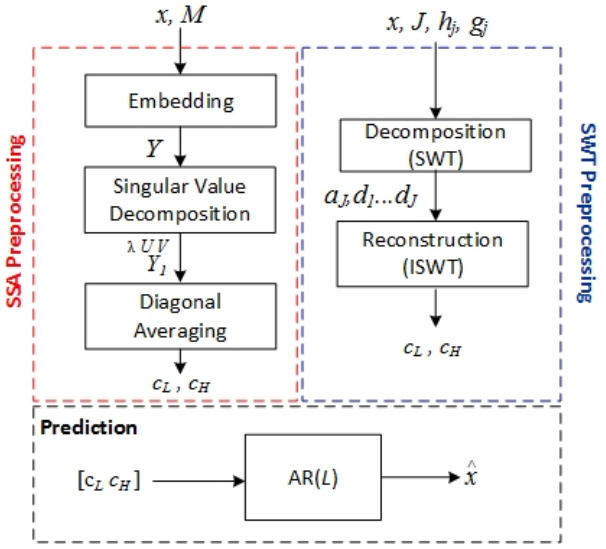

Fig. 1. Forecasting methodology

The one-step ahead forecasting model is defined with the next expression:

$$\hat{x}(n+1) = f\left[c_L(n), c_H(n)\right], \tag{1}$$

where $\hat{x}$ is the predicted value and n represents the time instant.

A. Preprocessing based on Singular Spectrum Analysis

The preprocessing through Singular Spectrum Analysis was illustrated in Fig. 1. The observed signal x through SSA is decomposed and the low frequency component c_L is extracted; while the component of high frequency c_H is computed by simple difference between the observed signal x and the component c_L. Conventionally the SSA implementation is defined in four steps: embedding, decomposition, grouping, and reconstruction by diagonal averaging [1]; but in this work the grouping step is not performed.

1) Embedding: The embedding step maps the time series x of length N in a sequence of M multidimensional lagged vectors of length K; the embedding process is shown below

$$Y = \begin{pmatrix} x_1 & x_2 & \dots & x_K \\ x_2 & x_3 & \dots & x_{K+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_M & x_{M+1} & \dots & x_N \end{pmatrix}, \tag{2}$$

where Y is a real matrix of MxK dimension, with $M < K$, and $K = N - M + 1$. The matrix Y is a Hankel matrix which means that the elements x_{ij} on the anti-diagonals $i + j$ are equal.

2) Decomposition: The decomposition step implements the SVD of the trajectory matrix Y. The SVD of an arbitrary nonzero $M \times K$ matrix $Y = [Y_1 : \dots : Y_K]$ is a decomposition of Y in the form

$$Y = \sum_{i=1}^{M} \sqrt{\lambda_i} U_i V_i^{\top}, \tag{3}$$

where λ_i is the ith eigenvalue of matrix $S = YY^{\top}$ arranged in decreasing order of magnitudes. $U_1, \dots, U_M$ is the corresponding orthonormal system of eigenvectors of the matrix S.
Standard SVD terminology calls $\sqrt{\lambda_i}$ the ith singular value of matrix Y; U_i is the ith left singular vector, and V_i is the ith right singular vector of Y. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called ith eigentriple of the SVD.

3) Reconstruction: The first eigentriple is used to extract the low frequency component c_L, the remainder eigentriples are not used, in that reason the grouping is not performed. The elemental matrix which contains the c_L component is computed with:

$$Y_1 = \sqrt{\lambda_1} U_1 V_1^{\top}. \tag{4}$$

The reconstruction is performed by diagonal averaging over Y_1. The elements $c_L(i)$ are extracted as follows:

$$c_L = \begin{cases} \frac{1}{k-1} \sum_{m=1}^{k} Y_1(m, k-m), & 2 \le k \le M, \\ \frac{1}{M} \sum_{m=1}^{M} Y_1(m, k-m), & M < k \le K+1, \\ \frac{1}{K+M-k+1} \sum_{m=k-K}^{M} Y_1(m, k-m), & K+2 \le k \le K+M. \end{cases} \tag{5}$$

The complementary component is the component of high frequency c_H, therefore $c_H = x - c_L$. Although c_H was not directly extracted by SSA, it was calculated from the component c_L, therefore c_H is an indirect product of the SSA decomposition.

ISSN 2395-8618

B. Stationary Wavelet Transform

The preprocessing through Stationary Wavelet Transform was illustrated in Figure 1. SWT is based on the Discrete Wavelet Transform, its implementation is defined in the algorithm of Shensa [22]. SWT implements up-sampled filtering [24, 25].

In SWT the length of the observed signal must be an integer multiple of 2^j, with $j = 1, 2, \ldots, J$; where J is the scale number. The signal is separated in approximation coefficients and detail coefficients at different scales, this hierarchical process is called multiresolution decomposition [26].

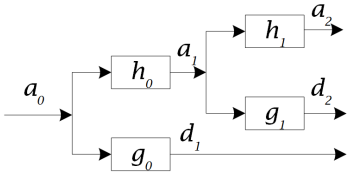

Fig. 2. Decomposition scheme of SWT (with 2 levels)

The observed signal a_0 (which was named x in previous section) is decomposed in approximation and detail coefficients through decomposition low pass filters $(h_0, h_1, \ldots, h_{J-1})$, and decomposition high pass filters $(g_0, g_1, \ldots, g_{J-1})$, one to each level as the scheme of Fig. 2. Each level filters are up-sampled versions of the previous ones.

At the first decomposition level, the observed signal a_0 is convoluted with the first low pass filter h_0 to obtain the first approximation coefficients a_1 and with the first high pass filter g_0 to obtain the first detail coefficients d_1. The process is defined as follows

$$a_1(n) = \sum_i h_0(i) a_0(n-i), \tag{6a}$$

$$d_1(n) = \sum_i g_0(i) a_0(n-i), \tag{6b}$$

This process follows iteratively, for $j = 1, \ldots, J$-1 and it is defined bellow:

$$a_{j+1}(n) = \sum_i h_j(i) a_j(n-i), \tag{7a}$$

$$d_{j+1}(n) = \sum_i g_j(i) a_j(n-i), \tag{7b}$$

Inverse Stationary Wavelet Transform (ISWT) performs the reconstruction. The implementation of ISWT consists in applying the operations that were done in SWT in inverse order and based on low-pass and high-pass reconstruction filters. The last coefficient approximation a_J reconstructs the component of low frequency c_L, whereas all detail coefficients reconstruct the component of high frequency c_H, both components are never decimated, therefore they have the same length as the observed signal.

III. Prediction Stage: Autoregressive model based on Components

The prediction is the second stage in the forecasting methodology (Fig. 1), and it depends of the first stage of preprocessing. The multistep-ahead forecasting is based on the direct method, which implements τ AR models of equal structure to predict the variable at time $n+h$ based on the linear relationship between L previous values of the components and the future value of the component at time $n+h$. The model uses the time series of length N which is split in two groups, training and testing, with length N_r and N_t respectively.

The following equation defines the general structure of the AR model in matrix notation:

$$\hat{y} = \beta Z, \tag{8}$$

where $\hat{y}$ is the predicted value of y with length N_r, Z is the regressor matrix and β is the linear coefficients matrix. The regressor matrix Z has order N_rx2L due to L lagged values of c_L and the L lagged values of c_H. The matrix of coefficients β has order 2x2L (one row for each component), and they are computed with the Least Square Method (LSM) [27], as follows:

$$\beta = Z^{\dagger} y, \tag{9}$$

where $Z^{\dagger}$ is the Moore-Penrose pseudoinverse matrix [27].

The direct strategy estimates τ models between regressors to compute h-step ahead prediction [28]. Each model returns a direct forecast of $\hat{x}(n+h)$. This strategy can be expressed as

$$\hat{x}(n+h) = f\left[Z(n), Z(n-1), \ldots, Z(n-L+1)\right], \tag{10}$$

where $h = 1, \ldots, \tau$.

IV. Efficiency Criteria

Two efficiency criteria are computed to evaluate the prediction accuracy of multi-step ahead prediction. One is the modified version of Nash-Suctlife Efficiency (*MNSE*); *MNSE* is computed to overcome the oversensitivity to extreme values induced by the mean square error of original Nash-Sutcliffe efficiency and to increase the sensitivity for lower values [29]:

$$MNSE = 1 - \frac{\sum_{i=1}^{N_t} |x_i - \hat{x}_i|}{\sum_{i=1}^{N_t} |x_i - \bar{x}|}, \tag{11}$$

where x_i is the ith observed value $\hat{x}_i$ is the ith predicted value, $\bar{x}$ is the mean of x, and N_t is the testing sample size.

The prediction accuracy was also evaluated through Mean Absolute Percentage Error (*MAPE*), Coefficient of Determination (R^2), and Relative Error (*RE*).

$$MAPE = \frac{1}{N_t} \sum_{i=1}^{N_t} \left| \frac{x_i - \hat{x}_i}{x_i} \right|. \tag{12}$$

$$R^2 = 1 - \frac{\sigma^2(x - \hat{x})}{\sigma^2(x)}, \tag{13}$$

ISSN 2395-8618

where σ^2 is the variance.

$$RE = \frac{(x - \hat{x})}{x}. \quad (14)$$

V. Experimental Results and Discussion

A. Data

The forecasting performance of hybrid models SSA-AR and SWT-AR are evaluated through a time series of injured in traffic accidents. The data were daily collected by the Chilean police and the National Traffic Safety Commission (CONASET) [23] from year 2000 to 2014 in Santiago, Chile with N =5479 records; the data reveals 260073 injured persons due to 58 causes defined by CONASET. In this work the problem is focused on principal causes via ranking; it was found that 15 causes are present in 80% of injured people in traffic accidents, which are categorized in *imprudent driving*, *pedestrian recklessness*, *signal disobedience*, *alcohol in driver*, *vehicle control loss*, and *mechanical causes*. The complex dynamic of the series is observed in Fig. 3. Nonstationary characteristics were probed with the KPSS test [30].

B. Prediction based on Singular Spectrum Analysis and the Autoregressive model

This forecasting is based on the component of low frequency extracted with SSA, and its complementary component of high frequency computed by subtraction. Therefore the performance model depends of the components that were provided. The order of the AR model was set by information of the Autocorrelation Function (ACF). Fig. 4a shows 7, 14, 21, 28, 31, and 35 day peaks; the multiple harmonic $L = 14$ was chosen to implement a parsimonious model.

One-step ahead forecasting was implemented with $L = 14$; the window length was evaluated through the metric *MNSE*, as it is shown in Fig. 4b. The effective window length was set in $M = 7$, which reaches a *MNSE* =98.4%; consequently the Hankel matrix has 7×5473 dimension. The component of low frequency c_L was extracted with SSA, and the component of high frequency is its complement. The SSA components are shown in Fig. 5. From Fig. 5, the c_L component presents slow fluctuation, whereas the c_H component presents quick fluctuation, the components quality is now evaluated in the prediction stage.

One step-ahead forecasting via SSA-AR is extended to multi-step ahead forecasting keeping the same settings ($L = 14$, $M = 7$), via direct strategy; Figures 6a, 6b and Table I present the results for multi-step ahead prediction of Injured in traffic accidents. From Figures 6a, 6b, and Table I, high accuracy was reached through the application of SSA-AR hybrid model for multi-step ahead prediction of injured in traffic accidents. SSA-AR presents a mean *MNSE* of 93.5% and a mean *MAPE* of 2.6%.

TABLE I
Multi-step ahead prediction, metrics *MNSE* and *MAPE*

	MNSE(%)		*MAPE*(%)	
h	SSA-AR	SWT-AR	SSA-AR	SWT-AR
1	98.4	99.8	0.6	0.07
2	96.7	99.4	1.3	0.2
3	94.8	98.8	2.1	0.5
4	93.1	97.6	2.7	1.0
5	91.5	95.1	3.4	2.0
6	90.2	87.4	3.9	5.2
7	89.4	82.5	4.3	7.1
Min	89.4	82.5	0.6	0.07
Max	98.4	99.8	4.2	7.1
Mean	93.5	94.4	2.6	2.3
Mean Gain		0.96%		13.0%

The observed signal vs the predicted signal for 7-days ahead prediction via SSA-AR is shown in Fig. 7a, good fitting is reached. The performance evaluation through *MNSE*, *MAPE*, R^2, and *RE* are presented in Table II. The SSA-AR model reaches high accuracy with a *MNSE* of 89.4%, a *MAPE* of 4.3%, a R^2 of 98.8%, and 97.4% of the predicted points show a relative error lower than $\pm 15\%$.

C. Prediction based on Stationary Wavelet Transform and the Autoregressive model

The wavelet decomposition was implemented with the mother wavelet function Daubechies of order 2 (Db2), in reason that Db2 presents better performance than other wavelets the time series decomposition with long term non-linear trend and periodic component [21]. The number of decomposition levels was evaluated in the prediction stage via AR(14) model. The decomposition level was set by evaluation of different values in the range $j = 1, \ldots, 4$ for one-step ahead prediction. As it is shown in Fig. 4c, the effective number of decomposition levels was set in $J = 2$, which reaches the highest efficiency *MNSE* of 98.8%. The last coefficient approximation reconstructs the component of low frequency c_L, whereas the addition of all detail coefficients reconstruct the component of high frequency c_H. The components extracted through SWT from the time series were illustrated in Fig. 5c and 5d. The c_L component present long-term fluctuations, whereas c_H present short-term fluctuations.

One step-ahead forecasting via SWT-AR is extended to multi-step ahead forecasting keeping the same settings ($L = 14$, $J = 2$), via direct strategy; Figures 6a, 6b and Table I present the results for multi-step ahead prediction of Injured in traffic accidents. From Figures 6a, 6b, and Table I, high accuracy was reached through the application of SWT-AR hybrid model for multi-step ahead prediction of injured in traffic accidents. SWT-AR presents a mean *MNSE* of 94.4% and a mean *MAPE* of 2.3%.

The observed signal vs the predicted signal for 7-days ahead prediction via SWT-AR is shown in Fig. 8a, good fitting is reached. The performance evaluation through *MNSE*, *MAPE*, R^2, and *RE* are presented in Table II. The SWT-AR model

ISSN 2395-8618

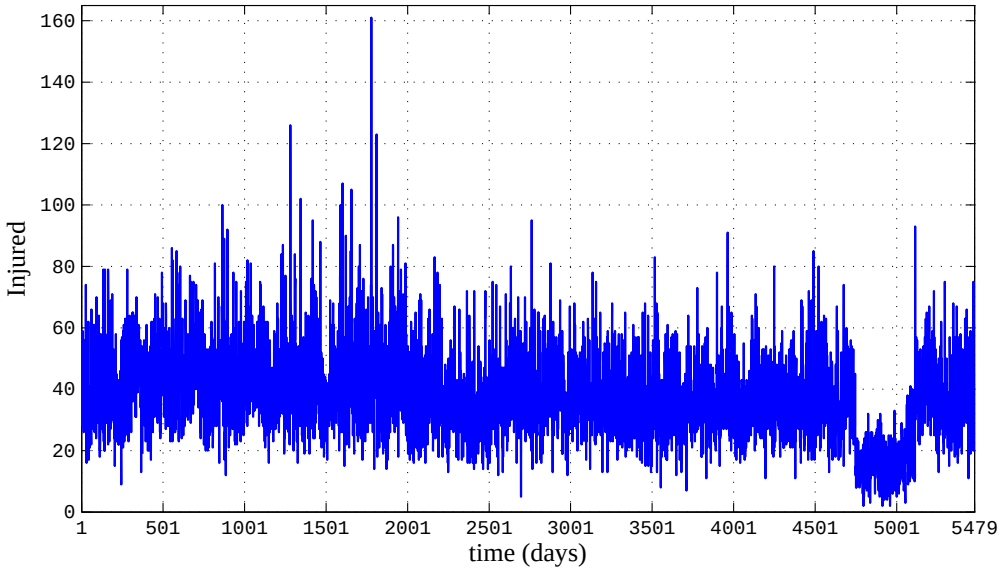

Fig. 3. Days vs Number of Injured by 15 principal causes

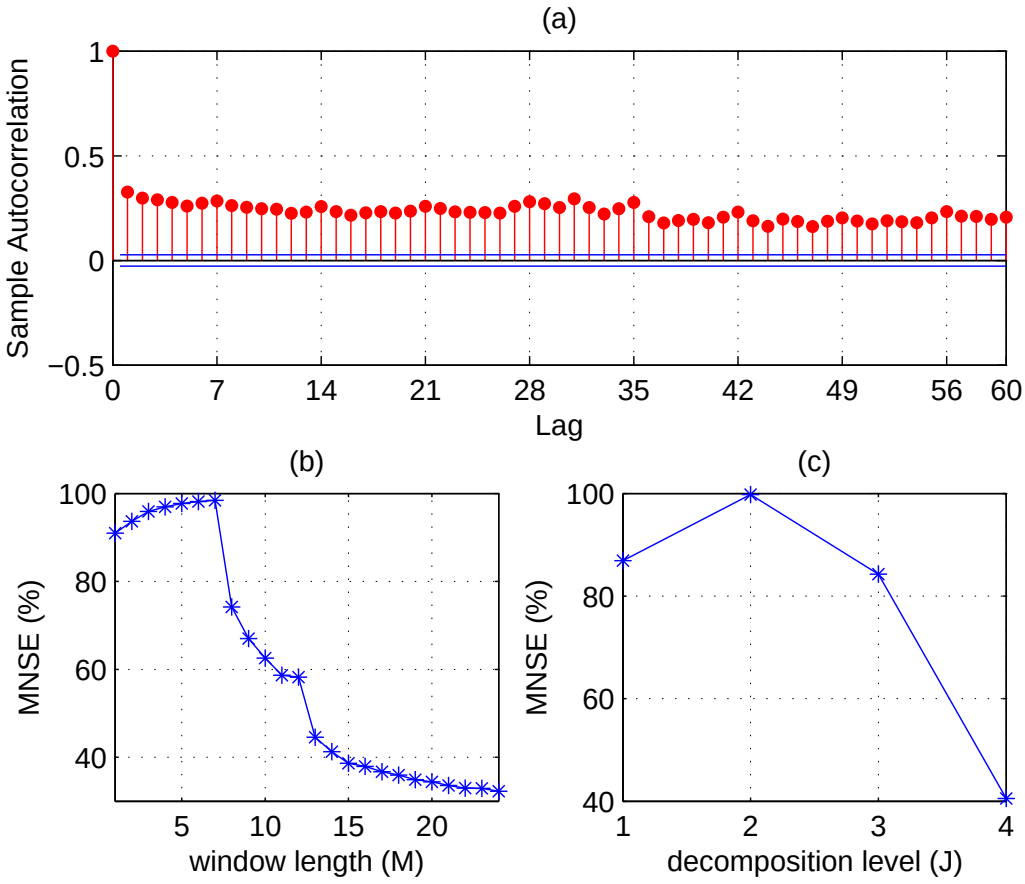

Fig. 4. (a) Injured in Traffic Accidents - ACF, (b) Evaluation of variable window length *M* for SSA, (c) Evaluation of variable decomposition level *J* for SWT

ISSN 2395-8618

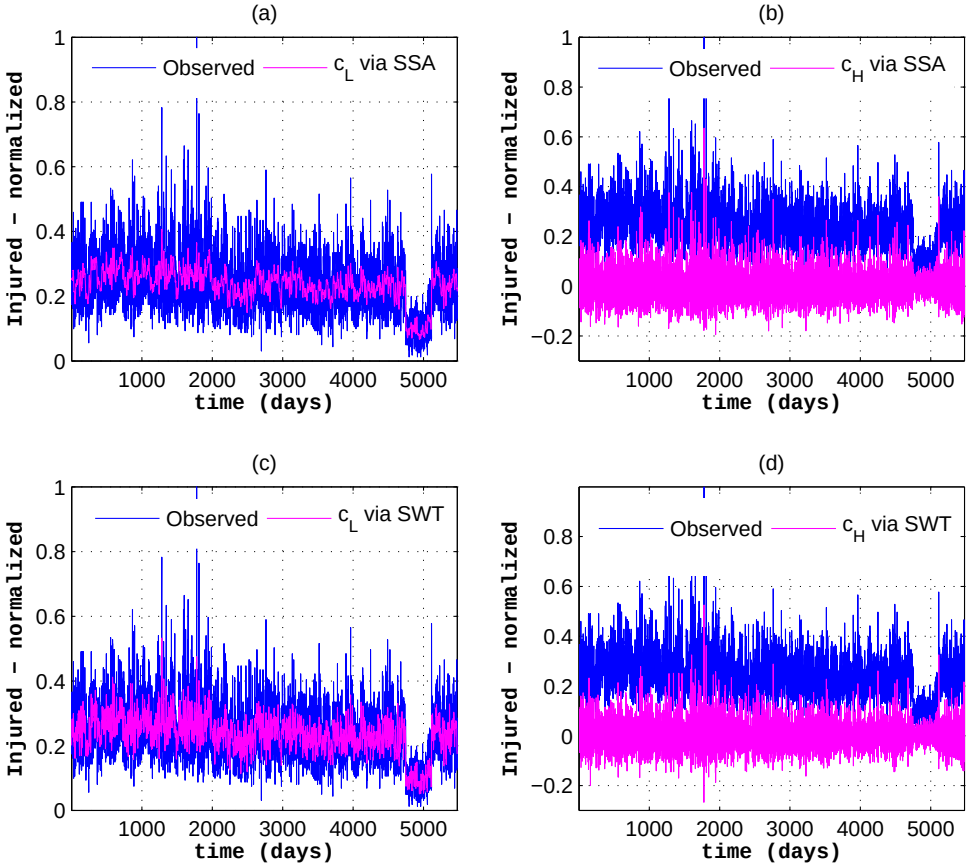

Fig. 5. Injured in traffic accidents Decomposition (a) c_L via SSA and SWT, (b) c_H via SSA and SWT

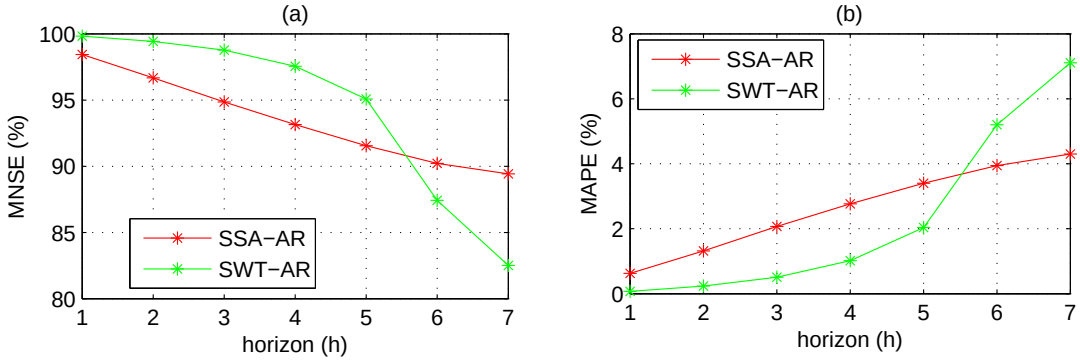

Fig. 6. Efficiency criteria for multi-step ahead prediction of Injured in traffic accidents, via SSA-AR and via SWT-AR (a) MNSE (b) MAPE

ISSN 2395-8618

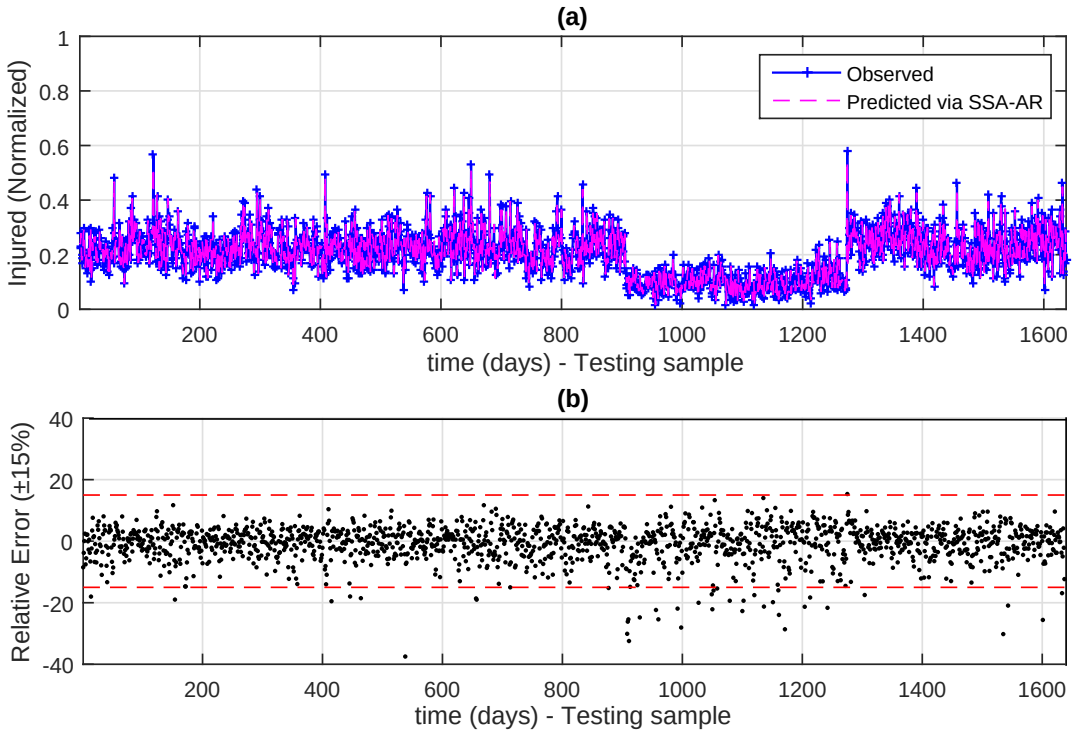

Fig. 7. Results for 7-day ahead prediction of Injured via SSA-AR (a) Observed signal vs Predicted signal, (b) Relative Error

reaches high accuracy with a *MNSE* of 82.5%, a *MAPE* of 7.1%, a R^2 of 96.5%, and 90.4% of the predicted points show a relative error lower than ±15%.

TABLE II
RESULTS FOR 7-DAYS AHEAD PREDICTION OF INJURED IN TRAFFIC ACCIDENTS

	MNSE(%)	*MAPE*(%)	R^2(%)	*RE*(%)
SSA-AR	89.4	4.3	98.8	97.4±15%
SWT-AR	82.5	7.1	96.5	90.4±15%
SSA-AR Gain	8.4%	65.1%	2.4%	7.7%±15%

From Table I, SWT-AR reaches a *MNSE* mean gain over SSA-AR of 0.96%, and a *MAPE* mean gain of 13.0%. However SSA-AR shown superiority with respect to SWT-AR for the farthest horizons ($h = 6$, $h = 7$) as it is presented in Fig. 6 and Table I. By instance from Table II, SSA-AR achieves a *MNSE* gain of 8.4%, a *MAPE* gain of 65.1%, a R^2 gain of 2.4% and 7.7% of *RE* gain (±15%) with respect to SWT-AR for 7-days ahead prediction.

VI. CONCLUSIONS

In this study have been presented and compared two hybrid prediction models based on Singular Spectrum Analysis and Stationary Wavelet Transform combined with the Autoregressive model. The models have been evaluated with a nonstationary time series daily collected from the traffic accidents domain in the period 2000 to 2014; the data characterize the fifteen most relevant causes of injured people in traffic accidents in Santiago, Chile.

The component obtained with the first eigentriple in SSA corresponds to the approximation signal obtained in the last decomposition level in SWT. On the other hand, the regarding eigentriples of SSA reconstruct the component of high frequency, which corresponds to the detail coefficients computed in SWT to reconstruct the component of high frequency. Both SSA and SWT, obtain components of low frequency with long-term fluctuations, and components of high frequency of short-term fluctuations.

The prediction results of SSA-AR and SWT-AR are similar in curve fitting and accuracy. SWT-AR shows the highest mean accuracy for multi-step ahead prediction with a mean *MNSE* gain of 0.96% and a mean *MAPE* gain of 13%. However SSA-AR achieves the best accuracy for farthest horizons; 7-days ahead prediction present a *MNSE* gain of 8.4%, a *MAPE* gain of 65.1%, a R^2 gain of 2.4%, and 7.7% (±15%) of *RE* gain.

Finally both models are suitable for traffic accidents forecasting, however further research can be undertaken to evaluate this potential techniques and the proposed strategies in the solution of other nonstationary problems.

REFERENCES

[1] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & Hall/CRC, 2001.

ISSN 2395-8618

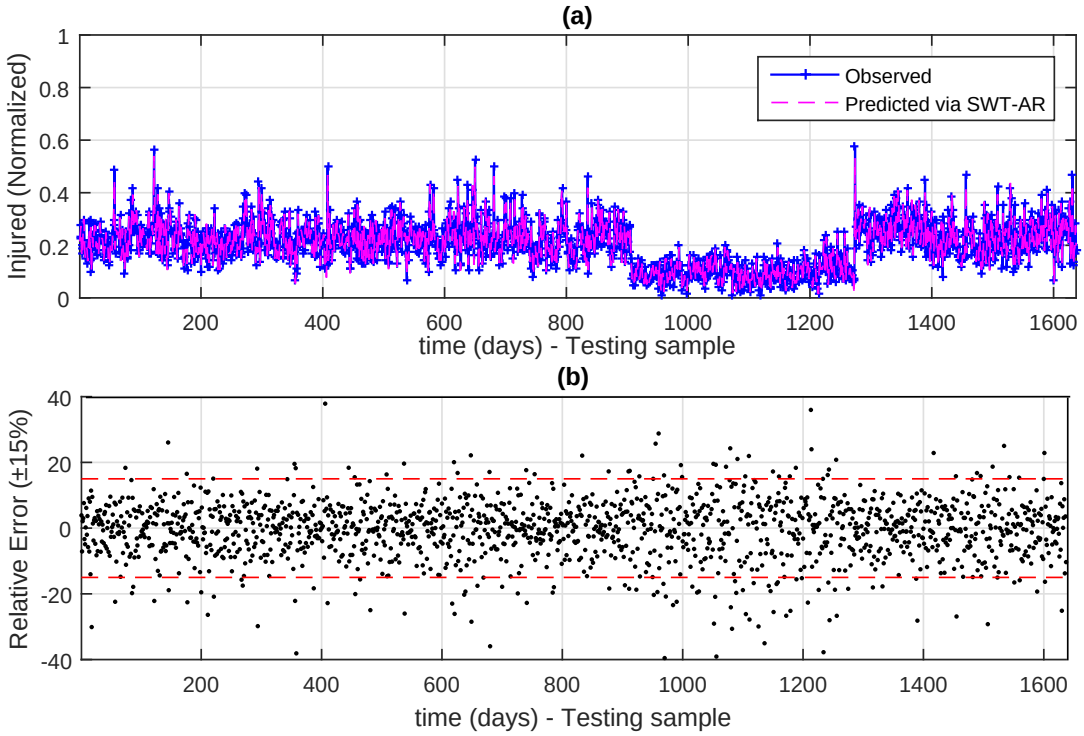

Fig. 8. Results for 7-day ahead prediction of Injured via SWT-AR (a) Observed signal vs Predicted signal, (b) Relative Error

[2] M. Loéve, "Sur les fonctions aleatoires stationnaires du second ordre," *Revue Scientifique*, vol. 83, pp. 297 – 303, 1945.

[3] K. Karhunen, "On the use of singular spectrum analysis for forecasting u.s. trade before, during and after the 2008 recession," *Annales Academiae Scientiarum Fennicae*, vol. 34, pp. 1 – 7, 1946, series A1, Mathematica-Physica.

[4] D. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data," *Physica D: Nonlinear Phenomena*, vol. 20, no. 2, pp. 217 – 236, 1986.

[5] R. Vautard, P. Yiou, and M. Ghil, "Singular-spectrum analysis: A toolkit for short, noisy chaotic signals," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1, pp. 95 – 126, 1992.

[6] M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, "Advanced spectral methods for climatic time series," *Reviews of Geophysics*, vol. 40, no. 1, pp. 3.1 – 3.41, 2002.

[7] U. Kumar and V. Jain, "Time series models (grey-markov, grey model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in india," *Energy*, vol. 35, no. 4, pp. 1709 – 1716, 2010, demand Response Resources: the US and International Experience.

[8] H. Hassani, S. Heravi, and A. Zhigljavsky, "Forecasting european industrial production with singular spectrum analysis," *International Journal of Forecasting*, vol. 25, no. 1, pp. 103 – 118, 2009.

[9] H. Hassani, A. Webster, E. S. Silva, and S. Heravi, "Forecasting u.s. tourist arrivals using optimal singular spectrum analysis," *Tourism Management*, vol. 46, no. 0, pp. 322 – 335, 2015.

[10] E. S. Silva and H. Hassani, "On the use of singular spectrum analysis for forecasting u.s. trade before, during and after the 2008 recession," *International Economics*, vol. 141, pp. 34 – 49, 2015.

[11] N. Brunsell, "A multiscale information theory approach to assess spatialtemporal variability of daily precipitation," *Journal of Hydrology*, vol. 385, no. 14, pp. 165 – 172, 2010.

[12] U. Okkan, "Wavelet neural network model for reservoir inflow prediction," *Scientia Iranica*, vol. 19, no. 6, pp. 1445 – 1455, 2012.

[13] R.-P. Liang, S.-Y. Huang, S.-P. Shi, X.-Y. Sun, S.-B. Suo, and J.-D. Qiu, "A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization," *Computers in Biology and Medicine*, vol. 42, no. 2, pp. 180 – 187, 2012.

[14] M. Protić, S. Shamshirband, D. Petković, A. Abbasi, M. L. M. Kiah, J. A. Unar, L. Zivković, and M. Raos, "Forecasting of consumers heat load in district heating systems using the support vector machine with a discrete wavelet transform algorithm," *Energy*, vol. 87, pp. 343 – 351, 2015.

ISSN 2395-8618

[15] S.-C. Huang, "Wavelet-based multi-resolution {GARCH} model for financial spillover effects," *Mathematics and Computers in Simulation*, vol. 81, no. 11, pp. 2529 – 2539, 2011.

[16] R. Jammazi and C. Aloui, "Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling," *Energy Economics*, vol. 34, no. 3, pp. 828 – 841, 2012.

[17] S. Jaipuria and S. Mahapatra, "An improved demand forecasting method to reduce bullwhip effect in supply chains," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2395 – 2408, 2014.

[18] A. Grossmann and J. Morlet, "Decomposition of hardy functions into square integrable wavelets of constant shape siam j," *Journal of Mathematics*, vol. 15, p. 723, 1984.

[19] A. Haar, "Zur theorie der orthogonalen funktionen-systeme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910.

[20] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 906–996, 1988.

[21] R. Maheswaran and R. Khosa, "Comparative study of different wavelets for hydrologic forecasting," *Computers & Geosciences*, vol. 46, pp. 284 – 295, 2012.

[22] M. J. Shensa, "The discrete wavelet transform: Wedding the trous and mallat algorithms," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[23] "Comisión Nacional de Seguridad de Tránsito," 2015. [Online]. Available: http://www.conaset.cl

[24] G. P. Nason and B. W. Silverman, *Wavelets and Statistics*. New York, NY: Springer New York, 1995, ch. The Stationary Wavelet Transform and some Statistical Applications, pp. 281–299.

[25] R. Coifman and D. Donoho, "Translation-invariant de-noising," pp. 125–150, 1995.

[26] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 7, pp. 674 – 693, 1989.

[27] G. H. Golub and C. F. V. Loan, *Matrix Computations*. The Johns Hopkins University Press, 1996.

[28] C. Hamzaebi, D. Akay, and F. Kutay, "Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3839 – 3844, 2009.

[29] P. Krause, D. P. Boyle, and F. B*ä*se, "Comparison of different efficiency criteria for hydrological model assessment," *Advances in Geosciences*, no. 5, pp. 89 – 97, 2005.

[30] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *Journal of Econometrics*, vol. 54, no. 1, pp. 159 – 178, 1992.

ISSN 2395-8618

Frequent Patterns Mining for the Satisfiability Problem

Celia HIRECHE, Habiba DRIAS, Neyla Cherifa BENHAMOUDA
USTHB, Department of Computer Science, LRIA, Algiers, Algeria
Email: chireche@usthb.dz, hdrias@usthb.dz, benhamoudaneyla@gmail.com
www.lria.usthb.dz

***Abstract*—This paper presents a novel approach for solving the Satisfiability problem by reducing its complexity. First, an improved, 'divide and conquer'version of the Apriori algorithm is introduced. It consists in dividing the problem instance into two or more if necessary, sub-instances and then in executing an ameliorated version of the Apriori algorithm for extracting the frequent variables appearing in the sub-instances.**
These most frequent variables are grouped into clusters and the corresponding problem are considered for resolution. Once done, the clusters can be shown as new smaller instances that are solvable separately using either the DPLL procedure or the BSO algorithm according to the number of variables to be solved.

***Index Terms*—Mining Frequent Patterns, Apriori, Meta-Apriori, Clustering, NP-Complete problems, Problem Solving, Satisfiability Problem, Complexity**

I. Introduction

ONE of the most important tasks of Data Mining[1] is the reducing complexity of data while keeping the integrity of the later.
Three kinds of treatment are used for this purpose, Classification and clustering which consist in dividing the data into small groups according to a certain training data for the classification, and according to the similarities between the elements for the clustering. The third process being the frequent patterns mining, which consists in extracting the most frequent items repeated together.
In this work, we aim at reducing the complexity of the satisfiability problem[2], the most known NP-Complete problem that arouses the most interest of the computational complexity community. The issue consists in finding an assignment to the variables to satisfy an instance represented as a CNF (Conjunctive Normal Form) Boolean formula.
There are two categories of solving approaches, the complete and the incomplete methods[3]. The first guarantees to find the optimal solution if it exists or proves that the problem cannot have a solution if appropriate. These methods cannot cope with large problem instances and would generate a combinatorial explosion and timeout calculation whatever the machine performance.
To get around these problems, the scientific community developed new methods based on approximation. These methods do not guarantee to find a solution even if it exists.
These constrains motivate us to think about a preprocessing -pretreatment- to execute before the resolution. This preprocessing step consists in using a frequent mining patterns to reduce the problem complexity by dividing it into sub problems (clusters) that can be solved separately in a second step, using a complete algorithm and an incomplete one.
The remainder of this document is organized as follows. The next section presents some interesting works related to the satisfiability problem and the Apriori algorithm. The satisfiability problem is then introduced in the third section. The fourth section is dedicated to the presentation of Bees Swarm Optimization Algorithm. The fifth section is consecrated to the Meta-Apriori algorithm, prior to presenting the Apriori-Clustering resolution in the sixth section. The conducted experiments and the obtained results are presented in the last section. Conclusions are finally summarized and some perspectives are suggested.

II. Related Works

Nowadays, several algorithms and solvers exist to get over the satisfiability problem, namely SAT.
The first category of SAT solvers deals with complete algorithms that are able to yield either a satisfying solution or a proof that such a solution does not exist. One of the most known and studied complete solver is the Davis-Putnam-Logemann-Loveland(DPLL)[4]. This backtracking algorithm recursively assigns a truth value to a variable and eliminate all clauses that contains it until being able to check whether the formula is satisfied. Several existing solvers are based on the DPLL algorithm.
An extension of the DPLL is introduced in the Conflict Driven Clauses Learning (CDCL)[3], in which a new clause is learnt when a conflict occurs while assigning values to the variables. Tens of CDCL based solvers exist nowadays[3].
In[5], the authors introduced the first parallel portfolio[6] SAT solver using a multicore architecture, allowing a communication between the four used cores (CDCL solvers) through lockless queues. These solvers are configured differently according to:

- the restart policy, using either a dynamic restart policy depending on the average size of the two last back-jumps, or an arithmetic one,
- the selecting heuristic, where they increase the random noise of the Variable State Independent Decaying Sum (VSIDS) heuristic to diversify the selecting process,
- the polarity using a progress saving politic, saving the polarity of variables between conflict and back-jump

ISSN 2395-8618

level, and a statistic polarity according to the occurrences of each literal (variable and its negation),
- the learning process, using the basic CDCL's implication graph[7] and introducing a novel arc, called inverse arc, that takes into consideration the satisfied clause,
- and finally, the clause sharing, which allows the communication between the cores.

All of these cores will deal with the whole base of clauses and try to solve it using different manner (configuration), which can be very time consuming. It would be a better choice to divide the problem so that the different solvers can cooperate by solving the different parts separately and save time.

The second category of solvers are based on incomplete algorithms. Their principle is to learn the problem's characteristics in order to guide the search without covering the whole search area.

One of the first and most studied incomplete algorithms is the Stochastic Local Search (SLS)[8], such as the famous GSAT. Firstly introduced in[9], the algorithm starts by assigning a truth value to all the variables, then generates the neighbourhood of the current solution by flipping the variables one by one. The choice of the variable to be flipped is made by selecting randomly an unsatisfied clause, and picking the variable that maximizes the number of newly satisfied clauses and minimize the number of newly unsatisfied clauses.

An extension of the GSAT was proposed in [10], named WALKSAT, in which the choice of the variable to be flipped is made by selecting from a random unsatisfied clause, the variable satisfying the GSAT condition with a probability p, and with a probability 1-p picks a variable randomly.Since then a family of WALKSAT solvers were created[11]. As other solvers based on SLS algorithm.

In[12], S. Cai et al. introduced a new two-mode SLS solver that combines be- tween two flip strategy. The first one being the CCA (Configuration Checking with Aspiration) heuristic, which does not allow flipping a variable if its configuration (neighbours) does not change since its last flip. And allow the flipping of those whose score is significant (their flip decreases the number of unsatisfied clauses significantly). If these kinds of variables do not exist, the flip strategy used is switched to the focused local search mode which selects a variable from a random unsatisfied clause. This solver has been combined with other solvers like glucose CCAnr+Glucose[13], which was presented in the SAT's competition 2014, and so others.

However, visiting the neighbourhood of each variable and counting the score of each variable at each step with the probability to switch to a random mode after this process is very time consuming.

In [14], the authors, being inspired by the Frankenstein's novel, introduced a solver which consists on a combination of existing high performance SLS SAT's techniques (solvers) with some mechanism that they introduced, using an automated construction process. The solver includes five parts or blocks, where the first is used for diversification -initializing selecting policy-. The three next parts are consecrated to the resolution itself; WALKSAT's based solvers for the second part, dynamic local search (penalties associated to the clauses) solvers for the third and GWSAT (joining GSAT-WALKSAT) for the fourth part. The fifth block is used to up to date data structures.

Even on this solver, selecting the solver to be used for resolution, can be very time consuming because of the diversity of problems instances.

Data Mining techniques were for the first time used for solving SAT in[15][16], where the authors used clustering[1] methods to reduce the problem instance into many groups using an intuitive method[15], creating a cluster for every new variable. A Genetic-K-Means where the clusters centres are generated using the genetic algorithm[1], and the classification is made using the K-means algorithm[1].

Many other Data Mining techniques exist, including Frequent Patterns Mining which consists on finding the patterns (items, variables, ...) that are repeated together. One of the most known algorithms used for mining frequent patterns is the Apriori Algorithm[1][17].

Introduced in[17], Apriori consists in finding the set of k-Itemsets that occur the most. A set of itemsets candidates is extracted to be validated by scanning the whole base which is very time consuming.

In[18], the authors considered the item with the minimum support, minimizing then the database scans and reducing the runtime. They also used the FP-growth algorithm in order to reduce the memory space.

III. THE SATISFIABILITY PROBLEM

Being one of the most studied NP-Complete problems, all eyes are turned to the Satisfiability problem, for its complexity and its impact on the whole NP- Completeness.

The Boolean Satisfiability problem[2], SAT, is to decide whether or not there is a satisfying assignment to the set of variables x making a Boolean formula (x) true. This formula being in conjunctive normal form (CNF) that is a conjunction of clauses, where each clause is a disjunction of literals, a literal being either a variable or its negation. In other words, find an assignment (true value to each variable) that satisfies all the clauses in the same time. Reminding that a clause is said to be satisfied (true) if and only if at least one of its variables is satisfied (true for a positive literal and false for a negative one).

The formal definition of the problem is shown in the following instance and question pair:

- Instance: m clauses over n literals
- Question: Is there any assignment of variables that satisfies all the clauses?

Example:

Let consider the following set of variables V = {X1, X2, X3} and the following set of clauses C = {C1, C2, C3, C4} defined as follow:

- C1 = X1, X2
- C2 = X2, -X3

ISSN 2395-8618

- C3 = -X1, -X3
- C4=X1, X2, X3

Note that the '-'means the negative form of the variable.

IV. BEES SWARM OPTIMIZATION FOR SOLVING SAT

The Bees Swarm Optimization Algorithm (BSO)[19] is a population-based search algorithm simulating the behaviour of bees when looking for food[20]. In fact, Karl Von Fris - 1946- observed that it is through a specified dance that the bees communicate the distance and the direction of the food source. The richest the source, the vigorous the dance so that when two sources of equal distance are found, the bees exploit the most potential area.

By analogy to the animals (reign), the BSO algorithm works as follow: First, an initial bee, named BeeInit generates a random solution named Sref, from which a search space namely SearchArea is determined using a diversification strategy. Each bee considers a solution of this SearchArea as a starting point to its local search and communicates the best solution found in a table called Dance. The best solution from this table is taken as the references solution (Sref) and the cycle restarts until no better solution to be found.

Algorithm 1 Bees Swarm Optimization Algorithm for SAT

```
Bees : table of bees
Solution : Variables ; Solution ; Evaluation
Sref ←Random Boolean Solution
while non stagnation do
    TL ←Sref
    SearchArea (Sref) : Random generation
    for ( i = 0 ←Bees count) : assign a solution of SearchArea to each bee do
        Local search
        Dance ←Best local search
    end for
    Sref ←Best solution from Dance using fitness procedure(attribute a point to the evaluation for each satisfied clause)
end while
```

V. META APRIORI

The Apriori[17][1] Algorithm is the most popular algorithm in data mining for extracting the frequent itemsets. It detects from a set of transactions, the items that are repeated the most together. It starts by extracting the singles frequent items to then recursively self-join the resulting itemsets until no longer itemset to be extracted (k-itemsets).

The Meta-Apriori algorithm[21] includes three steps; partitioning step, Apriori step, and fusion step. The partitioning step consists in dividing the database into two clusters or more if necessary. This partitioning is made by classifying the transactions according to the frequency of their items, having as result, almost the same items in both of the clusters with the same frequency. The Apriori step, as its name indicated, is the application of an improved Apriori algorithm on previous clusters. These improvements were introduced to reduce the Aprioris time consuming, and consist in:

- A vertical representation for a better representation of the database and the set of candidate itemsets, so that the entry of the structure is the item (vari- able) and the contents is the set of transactions (clauses) where it appears.
- Validation of a candidate when its frequency is equal to the support (the condition is satisfied).
- Elimination of the items that appear less than the minimum support, and the transactions containing a lesser number of item than the current itemset size.

To end with the fusion step, where the itemsets of both of the clusters (of all clusters if more than two) are joined.

Algorithm 2 Meta-Apriori Algorithm

```
Variables :
CS1, CS2 : sub-transaction base
Ci : ith itemsets candidates
Input :
TB: transaction base
MinSUp: minimum support
Output:
FPB: frequent patterns base
procedure DIVIDING(TB,CS1,CS2)
    for i :0 to TB length do
        CS1 ←TB[i] or CS2 ←TB[i] according to the frequency of the items on both of CS1, CS2
        Return CS1, CS2
    end for
end procedure
procedure APRIORI(TB,FBP)
    extract 1-itemset and validate
    while (itemset to be extracted ) do
        Ci= self join the itemsets (new candidates)
        Validation(Ci)
        i=i+1
    end while
    Return FBP
end procedure
```

VI. META-APRIORI CLUSTERING FOR SAT SOLVING

In this section, we propose a novel algorithm for solving the SAT problem, where data mining collaborates with a complete resolution algorithm and incomplete one.

With the aim of reducing the problems complexity, the problems instance is divided into two groups (clusters) using, as presented in the previous section, a frequency clustering, to then execute the improved Apriori algorithm, giving as result a set of k most frequent itemsets.

Two methods are then possible; the first method merges (fusion) the itemsets of the two instances, on one unique set of itemsets which is used for creating clusters by using these itemsets as cluster's centre (If two itemsets share more than half of the elements, the two centres are merged). The

ISSN 2395-8618

problem's instance is then classified -into these clusters- using the Hamming distance[22] , so that an itemset is classified in the cluster with which it shares the maximal number of items. These clusters can be seen as new problem's instances which can be solved either by using the DPLL algorithm or the BSO algorithm according to the number of variables to be solved. The resulted solutions are then merged. The second method follows a top-down schema. Contrarily to the first, it does not merge the itemsets of both instances groups but continues splitting the instances using the same process as that described in the first method (creating the clusters using the frequent itemsets, and classify the two instances separately). Once all clusters created, the resolution of each of the clusters is made using the DPLL and the BSO algorithms. The solutions obtained by all the clusters are then combined to yield the general problem's solution.
The following figure illustrates the two presented methods.

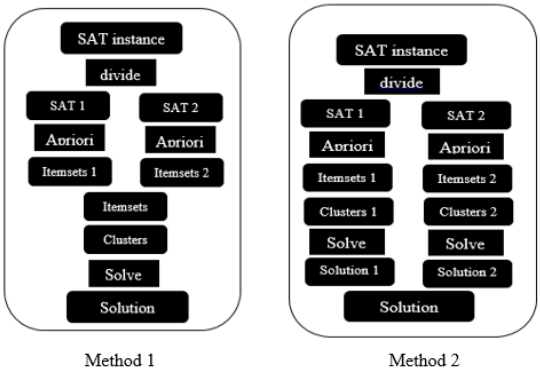

Fig. 1. META-APRIORI CLUSTERING FOR SAT SOLVING.

VII. EXPERIMENTS

To show the efficiency of the proposed approach, some experimentations were conducted on an i7 2.40 Ghz 4Go and the implementation on Microsoft visual studio CSharp 2013, and were conducted on some benchmarks which are presented in the Table 1.

TABLE I
BENCHMARKS DESCRIPTION

Benchmark	Solvability	Number of variables	Number of clauses
Benchmark1 [23]	Unsolved	99	8691
Benchmark 2 [23]	Solved	230	9975
Benchmark 3 [23]	Solved	440	9291
Benchmark 4 [23]	Solved	240	10409
Benchmark 5 [23]	Solved	260	11276
IBM 7 [24]	Solved	8710	39374
GALILEO 8 [24]	Solved	58074	276980
GALILEO 9 [24]	Solved	63624	307589

Table.1 describes the characteristics of each benchmark [23] [24], either they are solvable or not, the number of variables and clauses of each one. The sources from which these benchmarks were obtained are detailed in the references.

Table2 represents the solving rates and time solving of the Meta-Apriori Clustering DPLL-BSO vs the best time solver

TABLE II
SATISFACTION RATES AND SOLVING TIME FOR META-APRIORI CLUSTERING VS THE BEST BENCHMARK'S SOLVER.

Benchmark Name	Method	Rate (%)	Time (s)	Best Time Solver (s)	Best Solver
Benchmark1	1St Method	99, 61	37, 01		
	2Nd Method	99, 64	41, 35	-	-
Benchmark 2	1St Method	99, 10	1, 63	372, 14	Solver 1[23]
	2Nd Method	98,84	6,88		
Benchmark 3	1St Method	97,71	2,2	2088,76	Solver 2[23]
	2Nd Method	98,30	0,66		
Benchmark 4	1St Method	99,05	20,1	1257,86	Solver 3[23]
	2Nd Method	99,22	23,76		
Benchmark 5	1St Method	99,20	23,29	233,091	Solver 4[23]
	2Nd Method	99,27	27,66		

[23] in the corresponding SAT competition. It shows a significant difference between Meta-Apriori Clustering DPLL-BSO solving's time and the best time solving of each benchmark, which is due to Meta-Apriori Clustering that reduce significantly the complexity of the problem's instance, allowing an important time saving. However, we can see that the problem's instance is not 100% solved.
Comparing, for example, the 3rd benchmark for which the time solving of the best solver is about 2000s and the Meta-Apriori's time solving is about 2s which is 1000time less than the best solver even if the rate is about 97%. This rate can be handled by the use of a more efficient solver than pure DPLL and BSO.
The aim of this paper is the time saving by reducing the problem's complexity using Data Mining techniques.

TABLE III
SATISFACTION RATES AND TIME CONSUMING BETWEEN TWO CLUSTERING METHODS

	Meta-Apriori Clustering-DPLL-BSO Method 2		BSO-DM+DPLL	
Benchmark Name	Rate (%)	Time (s)	Rate (%)	Time (s)
IBM 7	199,11	13,98	93.77	24,25
GALILEO 8	99,06	617,97	97.65	1502,81
GALILEO 9	98,85	815,55	97.64	1620,47

Table 3, presents the rates and time consuming between the Meta-Apriori Clustering DPLL-BSO and the BSO-DM-DPLL[9]. These results, show that Meta-Apriori Clustering DPLL-BSO gives much better results (satisfiability rate) than the BSO-DM-BSO with time saving.

VIII. CONCLUSION

Throughout this paper, we proposed an approach based on mining frequent patterns associated with a complete algorithm and an incomplete one.
The proposed improvement of Apriori, Meta Apriori, is used as a preprocessing by extracting all the variables that appear together. The problem's instance being divided into clusters using the later itemsets, the problem's complexity is lesser, allowing the resolution of each of the clusters using either a complete algorithm or an incomplete one according to the number of variables to be solved. The later approach was applied to the Satisfiability problem because of its importance in the Artificial Intelligence community and the impact of

ISSN 2395-8618

solving such an important problem.

Many algorithms and solvers are proposed each year for solving SAT. The later approach was tested and the results of the experimentations show the impact of using frequent patterns mining as a preprocessing for solving problem.

We believe that this approach would be more efficient when used with a more efficient solver. For our future work, we will integrate this method in a solver that have proven his efficiency.

REFERENCES

[1] K. M. Han. J and P. J, "data mining, concepts and techniques," *Third Edition (The Morgan Kaufmann Series in Data Management Systems).*, 2011.

[2] J. Garey. M.R, "computers and intractability: A guide to the theory of np-completeness," *A Series of Books in the Mathematical Sciences. W. H. Freeman and Co.. pp. x+338. ISBN 0-7167-1045-5. MR 519066.*

[3] V. M. Biere.A, Heule. M and Wals.T, "Handbook of satisability ios press." *IOS press, 2009, ISBN 978-1-58603-929-5 (print)*, 2009.

[4] L. Davis. M and L. D, "a machine program for theorem proving," *Communications of the ACM 5(7)*, p. 394397, 1962.

[5] J. Hamadi. Y and S. L, "manysat: a parallel sat solver," *Journal of Satisfiability, Boolean Modeling and Computation 6(4)*, p. 245262, 2009.

[6] Gomes.C and S. B, "algorithm portfolios," *Artificial Intelligence,126(1-2)*, p. 4362, 2001.

[7] S. K. A. Marques-Silva. J.P, "grasp-a search algorithm for propositional satisfiability," *IEEE Transactions on Computers,48(5).*, vol. 232, p. 506521, May 1999.

[8] S. Hoos.H. H and Kaufmann.M, "stochastic local search : Foundations and applications," *Third Edition (The Morgan Kaufmann Series in Data Management Systems) Elsevier, ISBN: 1-55860-872-9..*, 2004.

[9] L. Selman. B and Mitchell.D.G, "a new method for solving hard satisfiability problems," *In: 10th AAAI, San Jose, CA*, p. 440446, 1992.

[10] K. Selman. B and C. B, ""local search strategies for satisfiability testing," *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 26, p. 521532, 1996.

[11] [Online]. Available: http://www.satlib.org/ubcsat/algorithms/

[12] L. C. Cai.S and S. k, "Ccanr: A configuration checking based local search solver for non-random satisfiability." *M. Heule and S. Weaver (Eds.)*, vol. 9340, p. 18, SAT 2015.

[13] L. Cai. S and Su.K, "Ccanr+glucose in sat competition 2014," *Proc. Of SAT Competition 2014*, vol. Solver and Benchmark Descriptions, p. 17, 2014.

[14] H. H. KhudaBukhsh.A.R, Xu. L and L.-B. K, "satenstein: Automatically building local search sat solvers from components," *Artificial Intelligence Journal (AIJ)*, vol. 232, pp. 20–42, March, 2016.

[15] D. H, H. C, and D. A, "datamining techniques and swarm intelligence for problem solving: Application to sat," *World Congress on Nature and Biologically Inspired Computing (NaBIC) 2013: IEEE ISBN: 978-1-4799-1414-2*, pp. 200–206, 2013a.

[16] D. A. Drias. H and H. C, "swarm intelligence with clustering for solving sat," *H. Yin et al. (Eds.): IDEAL 2013, LNCS 8206, Springer-Verlag Berlin Heidelberg*, p. 586594, 2013b.

[17] S. Agrawal.R, ""fast algorithms for mining association rules"," *1994 Int.Conf. Very Large Data Bases (VLDB94)*, vol. 99, p. 487499, Sept 1994,Santiago, Chile.

[18] G. A. Bhandari. A and D. D, "Improvised apriori algorithm using frequent pattern tree for real time applications in data mining." *In: Procedia Computer Science Vol. 46, International Conference on Information and Communication Technologies (ICICT 2015)*, vol. 46, p. 644651, 2015.

[19] S. S. Drias. H and Y. S, "a computing procedure for quantification theory," *In: Cabestany, J., Prieto, A.G., Sandoval, F. (eds.)IWANN 2005. LNCS,*, vol. vol. 3512, p. 318325, 2005.

[20] C. S. Seeley. T.D and S. J, "collective decision-making in honey bees: how colonies choose among nectar sources," *Behavioral Ecology and Sociobiology 28*, vol. 232, pp. 277–290, 1991.

[21] D. Benhamouda. N.C and Hireche.C, "Meta-apriori: a new algorithm for frequent pattern detection." *ACIIDS 2016, Part II, LNAI 9622*, vol. 99, p. 277285, 1994,Santiago, Chile.

[22] H. R., "error-detecting and error-correcting codes," *Bell System Technical Journal 29*, vol. 2, pp. 147–160, 1950.

[23] [Online]. Available: http://www.satcompetition.org/edacc/sc14/

[24] [Online]. Available: http://www.cs.ubc.ca/~hoos/SATLIB/benchm.html

ISSN 2395-8618

Analysis of the discrete wavelet coefficients using a Watermark Algorithm

Sandra L. Gomez Coronel, Marco A. Acevedo Mosqueda, Ma. Elena Acevedo Mosqueda and Ricardo Carreño Aguilera.

***Abstract*—This paper analyses the performance of the Discrete Wavelet Transforms (DWT) in a watermark algorithm designed for digital images. This algorithm employs a perceptive mask and a normalization process. The watermark insertion is done through the spread-spectrum technique, which is still, after a couple of decades, one of the safest ways to disguise the presence of the watermark in the digital image to the human eye. The algorithm is evaluated by establishing which wavelet coefficient provides the best accommodation in the watermark, i.e., it is not noticeable and will resist the various attacks, both intended and unintended. Different objective metrics are used-Peak Signal to Noise Ratio (PSNR), Multi-Scale Structural Similarity Index (MSSIM) average, correlation coefficient- and Bit Error Rate (BER) to determine which coefficient performs better in the insertion and extraction of the watermark.**

***Index Terms*—Discrete Wavelet Transform, Image normalization, Perceptive Mask, Spread Spectrum, Watermarking.**

I. Introduction

Nowadays most humans deal with information in a digital format (audio, video, or image). Although its immediate access represents an advantage, we cannot forget that the contents are also vulnerable to any kind of manipulation. By shielding digital data, we can safely share information, even through unsafe channels, preventing illegal reproductions or unauthorized alterations to original material. One way to achieve this is through watermarks, which purpose is to protect copyright in digital contents by inserting information into the digital file -that is to be authenticated. The watermark should remain imperceptible, robust, and hard to remove or alter; however, it must remain detectable when verifying the data. Over the last two decades, various watermarking techniques have been developed around three features: robustness, safety, and legibility. In the practice, the first two qualities work as opposites, because when imperceptibility is the focus, there is a tendency to loose robustness. When one intends to prevent visual alterations to the image, some of its perceptible areas remain unmodified, making the watermark vulnerable to both intentional and unintentional attacks. In addition, the legibility aspect seeks for the watermark to be detected, and/or extracted, at wish without any setbacks.

The different techniques found in the state-of-the-art demand the ability to insert the watermark in two levels: the spatial and the transform domains. The goal is to achieve an imperceptible watermark, impervious to all attacks because of its robustness. As a rule, the techniques suited for the spatial domain lack robustness, because the pixels (or pixel clusters) that must be marked are directly modified. To avoid perceptible changes, one option is to alter the least significant bit (LSB), or a cluster of them [1], [2], [3], [4], nevertheless, by modifying the intensity levels of the pixels, we end up with techniques that hold a small amount of robustness. This is the reason why it is preferable to work on the transform domain (Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), Contourlet Transform, or Hermit transform (HT)), thus making more difficult to eliminate or modify the watermark [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. There are also some techniques that take into account the features of the Human Vision System model (HVS) to hide the watermark [15], [16], [17], [18], [19]. These techniques are on the increase because of the positive results they produce in regards to intended and unintended attacks. Four approaches stand out: [16], [17], [20], [21]. The first one shows satisfactory results in JPEG compression and cropping, while successfully disguising the watermark through the DWT sub detail bands texture and luminance. [17] takes [16] as a frame of reference, but uses the Contourlet transform. Finally, in order to support more geometric attacks, algorithms like [20], [21] have resorted to normalized method of the marked image, preventing in this manner variations to affine transformations. Also there are other algorithms that proposed use Zernike moments or Scale-Invariant Feature Transform (SIFT) [22], [23], [24] to improve the selection places to insert the watermarking, ensuring robustness against attacks and quality image. Zernike moments have ability to provide faithful image representation and they are insensitivity to noise, whereas SIFT can extract feature points robust against various attacks, such as rotation, scaling, JPEG compression, and also transformation.

In light of the previous results, in this paper we suggest the evaluation of a watermark algorithm that uses a normalized process, as well as a perceptive watermark, so to guarantee robustness and prevent it to be perceptible. After dispersing the mark in the DWT domain, the watermark must be inserted in the spatial domain. The evaluation consists in establishing which is the best coefficient to disperse the watermark while obtaining the best results in relation to the robustness and quality in the marked image. Two aspects have then to be considered: on the one hand, even when the significant perceptual coefficients of the high-frequency subbands preserve

Sandra L. Gomez Coronel, Departamento de Ingeniería, Instituto Politécnico Nacional, UPIITA. Av. IPN No. 2580. Col. La Laguna Ticomán, Postcode 07340. Ciudad de México. México e-mail: (sgomezc@ipn.mx).

Marco. A. Acevedo-M, Ma. Elena Acevedo-M and Ricardo Carreño A. are with Instituto Politécnico Nacional, Sección de Estudios de Posgrado e Investigación, ESIME Zacatenco, Ciudad de México.

Manuscript received X; revised X.

ISSN 2395-8618

the invisibility of a watermark, these will remain vulnerable to common processing attacks; on the other, the low-frequency subbands coefficients cannot be modified, because such a change would be perceptible. Therefore, we suggest dispersing the watermark in the mid-frequency subbands coefficient (mid-low, and mid-high). We proposed to extract the watermarking not just its detection, because we use as watermarking legible information. Some algorithms use logos or pseudo random sequences, so the information amount is great. In this particular paper we use watermarks lengths between 60 and 104 bits. In order to allow the reader to comprehend the process, this paper presents the following structure: section two summarizes the DWT theory, while the third explains the proposal regarding the mark algorithm and watermark extraction; section four holds the results of each coefficient tests-to that end, several metrics were applied: Peak Signal-to-Noise Ratio *(PSNR)*, correlation coefficient, Multi-Scale Structural Similarity Index *(MSSIM)* average, and Bit Error Rate *(BER)*. The last section encloses the conclusions.

A. Discrete Wavelet Transform (DWT)

The wavelet transform can be understood as the decomposition of a group of basic functions, which can be obtained through scales and samplings of a mother wavelet. The analysis of this transform results in a group of wavelet coefficients that shows how close to a particular base function the signal actually is. Therefore, it is to be expected for any general signal to be represented as a decomposition of wavelets. This means that each original wave form can be synthesized through the constant addition of essential blocks that have different sizes and amplitude. Although there are many wavelet types, the Discrete Wavelet Transform (DWT) is the most common when processing images. The actual goal of the DWT is to convert a continuous signal into a discrete one through a sampling process. The latter is based on a multiresolution analysis, i.e. a specific number of decomposition levels in the wavelets domain. These are retrieved through a variety of digital filters (low-pass and high-pass filters).

1) Two-Dimensional Wavelet Transform: Digital images are two-dimensional digital signals, represented by a I matrix of *mxn* dimensions. The two-dimensional discrete wavelet transform requires [25]:

1) A scaling function $\varphi(x, y)$
2) Three two-dimensional wavelets $\psi^H(x, y)$, $\psi^V(x, y)$, $\psi^D(x, y)$

Each one is the product of the φ one-dimensional scaling function and the corresponding ψ wavelet, so that (Eq. 1) to (Eq. 4):

$$\varphi(x, y) = \varphi(x)\varphi(y) \tag{1}$$

Is a separable scaling function, and:

$$\psi^H(x, y) = \psi(x)\psi(y) \tag{2}$$

$$\psi^V(x, y) = \psi(x)\psi(y) \tag{3}$$

$$\psi^D(x, y) = \psi(x)\psi(y) \tag{4}$$

are separable wavelets.

These wavelets measure the intensity variations or gray levels. ψ^H measures the variations along the columns, that is, where the horizontal image's details are preserved, and the mid-low frequencies (*h* coefficient) held. ψ^V measures the variations along the rows, where the vertical details and mid-high frequencies (*v* coefficient) are enclosed. ψ^D measures the diagonal details as well as the high frequencies (*d* coefficient). The *a* coefficient holds the low frequencies and contains a compressed version of the original signal. The insertion of a watermark must occur in areas in which human vision is less sensible to changes, i.e. in the detail coefficients [26], [27]. Figure 1 shows a scheme of the two-dimensional wavelet decomposition, for a *x[n,m]* signal.

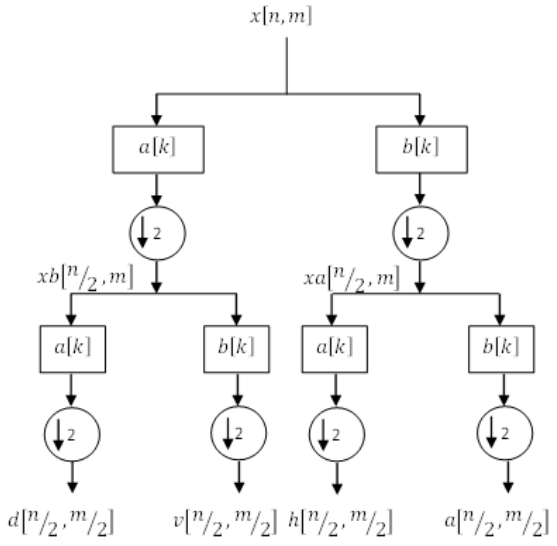

Fig. 1. Wavelet signal decomposition *x[n,m]*

II. Watermarking Algorithm

The purpose of this paper is to evaluate which of the wavelet coefficients is more suitable to disperse the watermark, by ensuring the marked image robustness and visual quality. Some approaches [26], [27], [28] have been set out so to establish which is the wavelet that guarantees better results based on the aforementioned parameters. This particular work focuses on to evaluate which wavelet coefficient produces the best results by inserting a watermark. The suggested algorithm uses a normalized method [20] to avoid alterations in the marked image due to possible geometric transformations. It also utilizes a perceptive mask that allows for the watermark to remain hidden, in the chosen coefficient, to the human eye. Each process is explained in the next sections.

A. Image normalization

The normalized process is based on the invariant moments theory [29]. For a *f(x,y)* image with *MxN* dimensions, these steps mus be followed.

ISSN 2395-8618

1) The $f(x, y)$ image mus be translated into $f_1(x, y) = f(x_a, y_a)$, with a center equivalent to the central mass of $f(x, y)$, and is given by (Eq. 5):

$$\begin{pmatrix} x_a \\ y_a \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix} - d \tag{5}$$

where (Eq. 6) and (Eq. 7) are:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{6}$$

$$d = \begin{pmatrix} d_x \\ d_y \end{pmatrix} \tag{7}$$

The values d_x, d_y are given by the geometric moments (Eq. 8):

$$d_x = \frac{m_{10}}{m_{00}}, d_y = \frac{m_{01}}{m_{00}} \tag{8}$$

where (Eq. 9):

$$m_{pq} = \left[\sum_{x=0}^{M-1} \sum_{j=0}^{N-1} x^p y^q f(x, y) \right] \tag{9}$$

2) Next, a shearing transform is applied in X direction to the $f_1(x, y)$ image, to get $f_2(x, y)$ using (Eq. 10):

$$A = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix} \tag{10}$$

where β is determined by (Eq. 11):

$$\mu_{30} + 3\beta^3 \mu_{12} + \beta^3 \mu_{30} \tag{11}$$

and μ_{pq} are the image's central moments.

3) A shearing transform is apply in Y direction to the $f_2(x, y)$ function, to get $f_3(x, y)$ with the matrix (Eq. 12):

$$A = \begin{pmatrix} 1 & 0 \\ \gamma & 1 \end{pmatrix} \tag{12}$$

where (Eq. 13):

$$\gamma = \frac{\mu_{11}}{\mu_{20}} \tag{13}$$

where μ_{pq} are the central moments of image resulting of step 2.

4) The $f_3(x, y)$ image is scale in both directions (x, y) to get $f_4(x, y)$, with the matrix (Eq. 14):

$$A = \begin{pmatrix} \alpha & 0 \\ 0 & \delta \end{pmatrix} \tag{14}$$

where α and δ are determined by the sized for the image obtained in the previous step.

The $f_4(x, y)$ image is the normalized image of the original $f(x, y)$ image, so that the watermark can be built as a function of the invariance, and becomes robust against different manipulations.

B. Perceptive Mask

In order to achieve an imperceptible watermark in the image to which it is inserted, it must be hidden through a mask. We call masking to the phenomenon by which a signal's visibility diminishes in favor of another one that disguises the original image. The design of the perceptive mask uses the human visual system model (HVS). Some works [14], [16], [17] take into consideration the texture and the luminance contents of the image subbands. Here, however, the perceptive mask is designed according to the Schouten brightness model [30]. It establishes that the brightness representation is invariant to the properties of a luminous source, as well as to the observation conditions. Watson [31], on the contrary, suggested designing the mask through a quantization matrix that depended on the image, thus producing a minimal erroneous bits rate for a given perception error, and vice versa. Originally, the Discrete Cosine Transform (DCT) was used, but the algorithm here described employs the HT. This adjustment was suggested in [32]. The decision to work with the HT responds to it's properties, as well as to the existing similarities between the functions of the synthesis filters, and those that model the receptive fields of the HSV. In [32], the contrast is calculated through the Hermite coefficients, and through the luminance masking. Eq. 15, Eq. 16 and Eq. 17 demonstrate the calculations pertaining to each one of the elements.

$$C = \left[\sum_{i=1}^{m} \sum_{j=1}^{n-m} C_{i,j}^2 \right]^{\frac{1}{2}} \tag{15}$$

where $C_{i,j}$ represents the Hermite Transform Cartesian Coefficients.

$$C_{thr} = k_0 \left(C_{min} + \left| \frac{B^\alpha - L_{min}^\alpha}{B^\alpha + L_{min}^\alpha} \right|^{\frac{1}{\alpha}} \right) \tag{16}$$

where:

k_0, is a constant.

C_{min}, is the minimal contrast present when a luminance level exists.

L_{min}, represents the maximum contrast sensitivity [32].

α, is a constant that includes values in the $[0, 1]$ interval.

C_{thr}, is the contrast masking.

B, is the brightness map proposed by [30].

and

$$M = k_1 max \left(C_{thr}, C^\beta C_{thr}^{1-\beta} \right) \tag{17}$$

where:

k_1 is a constant.

M is the perceptive mask.

C. Watermark Insertion

The watermark insertion process is illustrated in figure 2, and includes the following steps:

1) Calculate the normalization parameters of the original image *I(n×m)*, to obtained $I_N(\hat{n} \times \hat{m})$.

ISSN 2395-8618

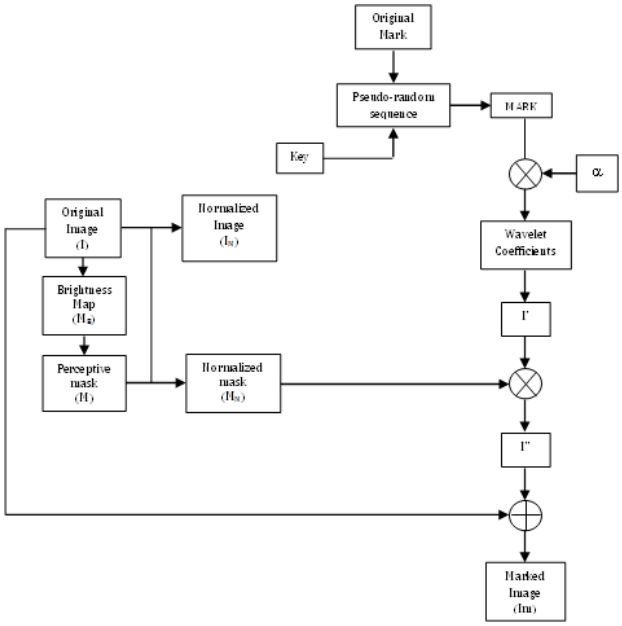

Fig. 2. Watermark insertion process based on the suggested *x[n,m]* algorithm

2) Create the binary watermark with a $n\{0,1\}$ length, starting from a numeric or alphanumeric code.
3) Generate p_i pseudo-random sequences, using a private key k, where $i = 1, ..., l$ and l represents the number of message bits applied as the watermark. Each sequence has $\{-1, 1\}$ values and $\hat{n} \times \hat{m}$ dimensions.
4) Calculate the brightness map B of the original image $I(n \times m)$ [30].
5) Calculate the perceptive mask M according to (Eq. 17)
6) To obtain M_N, normalize the perceptive mask M using the normalization parameters resulting after step 1.
7) Modulate the watermark with the p_i sequences to obtain W_a (Eq. 18)

$$W_\alpha = \sum_{i=1}^{l} (2m_i - 1)p_i \tag{18}$$

where m_i is the *i-th* bit of the watermark.
8) Generate the null wavelet coefficients and choose those in which the watermark will be inserted.
9) Insert the watermark through (Eq. 19):

$$\tilde{I_{k,l}}(i,j) = \alpha W_a \tag{19}$$

where:
α is a strength control parameter.
W_a is the modulated watermark.
$I_{k,l}$ is the modified wavelet coefficient.
10) Calculate the inverse wavelet transform of the coefficients to get $\hat{I}$.
11) Multiply $\hat{I}$ with the normalization mask M_N and apply the inverse normalization process to get $\hat{\hat{I}}$.
12) The final watermark W is inserted in the original image in additive form through (Eq. 20):

$$I_m = I + \hat{\hat{I}} \tag{20}$$

D. Watermark Extraction

To extract the watermark a correlated detector is to be used during the process, so that, when comparing the resulting correlation value of the sample with the original, one must consider if it is a bit 1 or a bit 0.

III. Test Results

We used 36 different images of dimensions 512×512 as well as two watermarks with 64 bits and 104 bits in length, respectively. The goal was to determine which wavelet coefficient, *h* y *v* , was more suitable to insert the watermark. Likewise, the strength control parameter α was modified (0.05 to 0.14 with increases of 0.1) in order to establish the value that throws the best results regarding the quality of the marked image, and the mark extraction. Since one of the purposes was to obtain a broad view, various types of metrics were used: Peak to Signal-to-Noise Ratio *(PSNR)*, that measures the statistical variations present between the original and the watermarked image, the Multi-Scale Structural Similarity Index *(MSSIM)* average and the coefficient correlation. In addition, the Bit Error Rate *(BER)*allowed the calculation of the modified bits quantity existing in each inserted mark. The averages of every metric helped us to compare the coefficients. Figures 3 to 7 show coefficient averages for both watermark lengths.

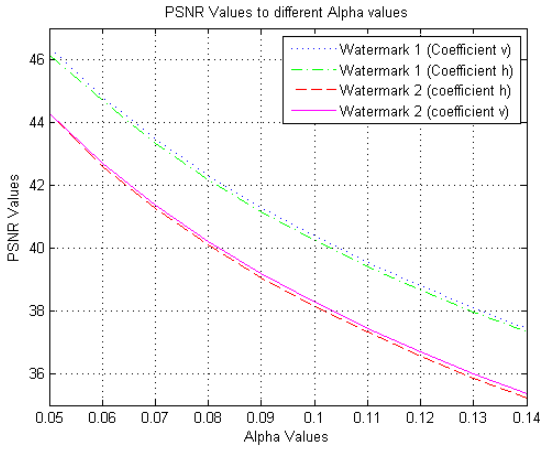

Fig. 3. *PSNR* average for each coefficient (*h* and *v*), after the insertion of both watermarks.

Figures 6 and 7 illustrate the watermark extraction. To make the insertion in the *v* coefficient entails more modified bits during the extraction process, thus obtaining a bigger *BER* in that specific coefficient. For the 64 bits length watermark, the erroneous bits average maintains up to a 3 bits average during the extraction, whereas the *h* coefficient has a 2 bit average. Now, when dealing with 104 bits long watermark, we face a similar situation: the best extraction results are related to the coefficient *h*-the modified bits average is of 4 bits-, while in the coefficient *v* are near to 6 bits. We concluded that to achieve a lower error average during the extraction, coefficient *h* is better to insert the watermark. Figures 8 and 9 show both the original and the modified Lena image using the coefficient *h*.

ISSN 2395-8618

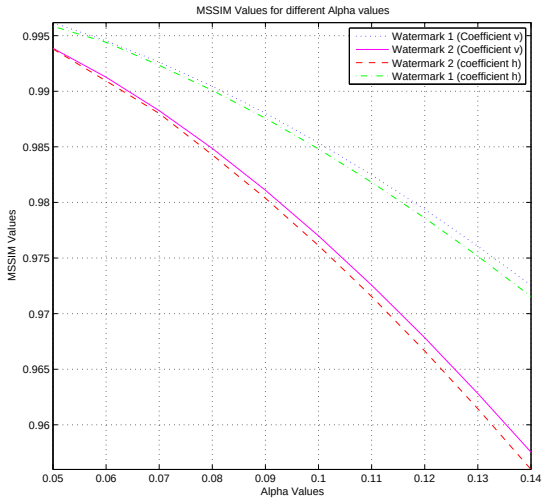

Fig. 4. *MSSIM* for each coefficient (h and v), after the insertion of both watermarks.

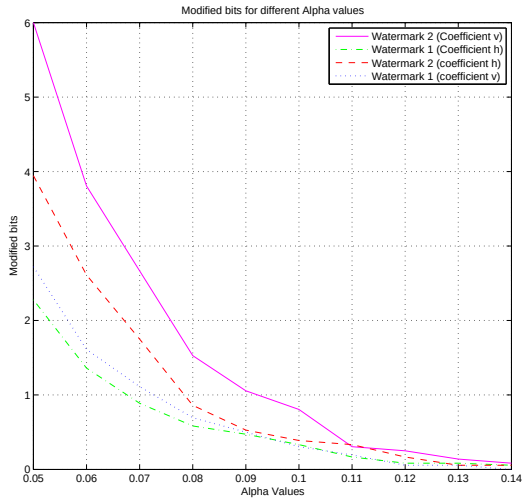

Fig. 6. Each coefficient (h and v) average, after the insertion of both watermarks.

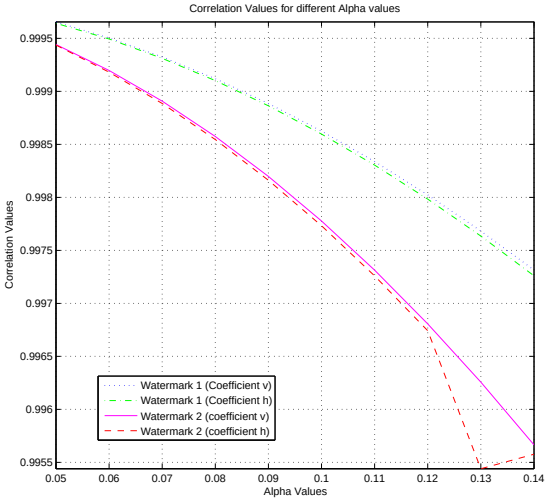

Fig. 5. Correlation coefficient average (h and v), after the insertion of both watermarks.

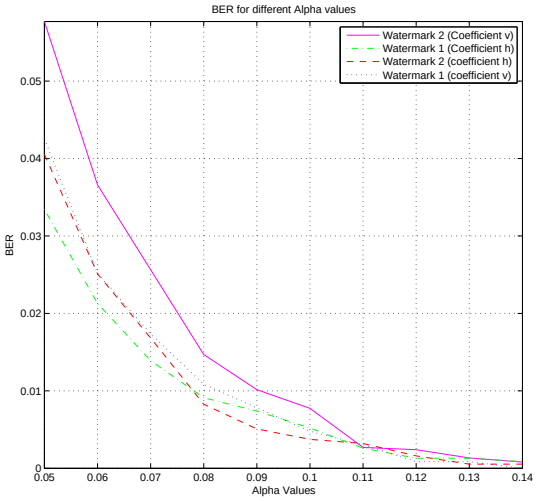

Fig. 7. BER average for each coefficient (h and v), after the insertion of both watermarks

.

A. *Evaluation of the Algorithm Attacks*

In order to determine if a similar result was obtained with attacks-after the watermark insertion in both h or v coefficients- three different types were tested: Gaussian filter, and Shearing in horizontal and vertical orientations. Concerning the Gaussian filter, a size $N \times N$, linear filtering was used; the filter average was 0 and the standard deviation was 0.5. Both parameters remained constant in all the tests. The only alteration was the N filter size-from 1 to 9, in 1 increments. Now, in case of shearing, in both cases X, Y, deformation was applied from 0 to 1 in 0.04 increments, which resulted in 26 distortions in each instance. These attacks were applied to demonstrate the performance during common processing and geometric attacks. Table I illustrates a representative sample of the failed attacks on 7 different images (these are commonly utilized in this type of tests). Each column indicates the total figure of failed attacks with an extraction of 2 modified bits at least. Despite of it, is it possible to recognize watermark.

As shown in table I it appears to be meaningless which coefficient is used to insert the watermark, since most attacks are unsuccessful. However, it is important to stress that the watermark extraction works better when the coefficient h is used. Therefore, we concluded that the latter is the best option when inserting a watermark, because it will accomplish both robustness and quality in the marked image. Finally, it is a fact that, for these sample images, we have a robust algorithm against to common processing and geometric attacks.

Figures 10, 11 and 12 show Lena image after all of the attacks that hold the highest parameters. Each one extracted perfectly the inserted watermark in the coefficient h.

ISSN 2395-8618

Fig. 8. Original image Lena

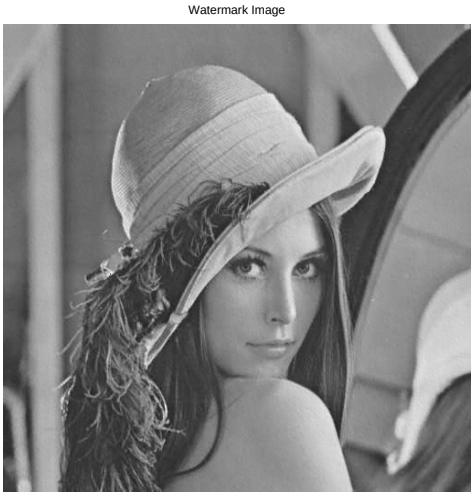

Fig. 9. Watermarked image Lena

Fig. 10. Lena image after horizontal shearing

TABLE I
FAILED ATTACKS FOR EACH IMAGE TESTED.

Image	Coeff.	Watermark	G. Filter	Shearing X	Shearing Y
Lena	h	1	9	22	8
		2	9	26	14
	v	1	5	24	23
		2	8	24	21
Babbon	h	1	5	26	22
		2	5	26	23
	v	1	3	4	26
		2	5	3	25
Barbara	h	1	5	26	23
		2	5	26	23
	v	1	5	24	21
		2	5	22	26
Boat	h	1	9	25	21
		2	9	26	22
	v	1	9	22	8
		2	9	24	8
Peppers	h	1	5	24	18
		2	1	21	21
	v	1	5	8	10
		2	9	6	9
Pirate (actor)	h	1	9	3	20
		2	9	0	18
	v	1	5	26	17
		2	5	26	21
Bridge	h	1	5	26	26
		2	5	26	26
	v	1	5	10	24
		2	9	14	25

Fig. 11. Lena image after vertical shearing

ISSN 2395-8618

Fig. 12. Lena image after Gaussian filter

IV. Conclusion

This paper presents the evaluation of a robust watermarking technique in order to determine the most suitable wavelet coefficient (h o v) in which to insert a l length watermark. According to the tests, we can firmly conclude that the coefficient h shows the best performance in regards to the insertion and extraction of the mark, as well in relation to resisting attacks. The values of the *PSNR* averages are close to $40dB$ even when the insertion force is altered. Such values indicate that the human eye is incapable of registering any difference in the marked image [30], [31]. Now, the remaining metrics (*MSSIM* and correlation coefficient) show averages closer to the unit, which means that, even when an image suffers alterations, they will stay hidden when compared to the original. The marked Lena image (Figure 9) shows that, visually, there are no noticeable changes when compared to the original image. As noted, one of the parameters that must be taken into account in the algorithm design is the robustness, because it is usually exposed to both unintended and intended attacks [34], the latter have more relevance because they specifically seek to affect the watermark. With this in mind, the algorithm, through a representative sample, was evaluated through a geometric attack that distorts the horizontal and vertical planes, and a common processing attack applying Gaussian filter, the results show that the coefficient h allows a better extraction of the watermark. Table I helps us conclude that, even when it is possible to make an extraction with both coefficients, more extractions are likely to occur in the various attacks when using the coefficient h. If we also add the metrics employed to measure the quality of the marked image, the values remain close to the ideal. Hence, the h coefficient is where the watermark must be placed.

Acknowledgment

The authors would like to thank for financial support at Instituto Politécnico Nacional IPN (COFAA, EDI and SIP), CONACyt-SNI.

References

[1] S. Kimpan and A. Lasakul and S. Chitwong. Variable block size based adaptive watermarking in spatial domain. *IEEE International Symposium on Communications and Information Technology, Vol. 1*, pages 374–377. 2004.

[2] G. Voyatzis and I. Pitas. Applications of toral automorphism in image watermarking *IEEE International Conference on Image Processing, Vol. 2*, pages 237–240. 1996.

[3] Ch-Ch. Chang and J-Y. Hsiao, Ch-L. Chiang. An Image Copyright Protection Scheme Based on Torus Automorphism *Proceedings of the First International Symposium on Cyber Worlds*, pages 217–224. 2002.

[4] I. J. Cox and J. Kilian, F. T. Leighton and T. Shamoon. Secure spread spectrum watermarking for multimedia *IEEE Transactions on Image Processing, Vol. 6, No. 12*, pages 1673 –1687. 1996.

[5] S.D. Lin and Ch-F. Chen. A robust DCT-based watermarking for copyright protection *IEEE Transactions on Consumer Electronics, Vol. 46, No. 3*, pages 415–421. 2000.

[6] C. M. Kung and J. H. Jeng and T. K. Truong. Watermark technique using frequency domain *14th International Conference on Digital Signal Processing, Vol. 2*, pages 729–731. 2002.

[7] H. Zhou and Ch. Qi and X. Gao. Low luminance smooth blocks based watermarking scheme in DCT domain *International Conference on communications, circuits and systems proceedings, Vol. 1*, pages 9–23. 2006.

[8] R. Dugad and K. Ratakonda and N. Ahuja. A new wavelet-based scheme for watermarking images *International Conference on Image Processing, Vol. 2*, pages 419–423. 1998.

[9] Z. Dawei, Ch. Guanrong and L. Wenbo. A chaos-based robust wavelet-domain watermarking algorithm *Chaos, Solitons, & Fractals, Vol. 22, No. 1*, pages 47–54. 2004.

[10] H. Dehghan and S. Ebrahim Safav. Robust Image Watermarking in the Wavelet Domain for Copyright Protection *CoRR", Vol. abs/1001.0282.* 2010.

[11] Ch-Ch. Chang and K-N. Chen and M-H. Hsieh. A robust public watermarking scheme based on DWT *Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 21–26. 2010

[12] E. Candés and L. Demanet and D. Donoho and L. Ying. Fast discrete curvelet transforms *Multiscale Modeling & Simulation, Vol. 5, No. 5*, pages 861–899. 2005.

[13] M. Jayalakshmi and S. N. Merchant and U. B. Desai. Blind Watermarking in Contourlet Domain with Improved Detection *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 449–452. 2006.

[14] N. Baaziz, B. Escalante-Ramírez and O. Romero. Image Watermarking in the Hermite Transform Domain with Resistance to Geometric Distortions *Proceedings of the SPIE Optical and Digital Image Processing*, pages 70000X1–70000X11. 2008.

[15] R. B. Wolfgang, C. I. Podilchuck and E. J. Delp. Perceptual watermarks for digital images and video *Proceedings of the IEEE, Vol. 87, No. 7*, pages 1108–1126. 1999.

[16] M. Barni, F. Bartolini and A. Piva. Improved wavelet-based watermarking through pixel-wise masking, *IEEE Transactions on Image Processing, Vol. 10, No. 5*, pages 783–791. 2001.

[17] N. Baaziz. Adaptive watermarking schemes based on a redundant contourlet transform, *IEEE International Conference on Image Processing, Vol. 1*, pages I221–4. 2005.

[18] A. Saadane. Watermark strength determination based on a new contrast masking model *Proceedings of the SPIE-IS& T Electronic Imaging, Vol.5020*, pages 107–114. 2003.

[19] S. Marano, F. Battisti, A. Vaccari, G. Boato and M. Carli. Perceptual data hiding exploiting between-coefficient contrast masking *Image Processing: Algorithms and Systems VI, Vol. 618*, pages 68121E1–68121E10. 2008.

[20] P. Dong, J. G. Brankov, N. P. Galatsanos, Y. Yang and F. Davoine. Digital watermarking robust to geometric distortions *IEEE Transactions on Image Processing, Vol. 14*, pages 2140–2150. 2005.

[21] M. Cedillo and M. Nakano and H. Pérez. Robust Watermarking to Geometric Distortion Based on Image Normalization and Texture Classification *51st Midwest Symposium on Circuits and Systems*, pages 245–2489. 2018.

[22] X-Ch. Yuan and Ch.M. Pun. Feature extraction and local Zernike moments based geometric invariant watermarking *Multimedia Tools and Applications, Vol. 72, No.1*, pages 777–799. 2014.

[23] C. Singh and S. K. Ranade. Image adaptive and high capacity watermarking system using accurate Zernike moments *IET Image Processing, Vol. 8, No. 7*, pages 373–382. 2014.

ISSN 2395-8618

[24] J. Xu. and L. Feng. A feature-based robust digital image watermarking scheme using image normalization and quantization *2nd International Symposium on Intelligence Information Processing and Trusted Computing (IPTC*, pages 67–70. 2011.
[25] I. Orea. Thesis: Marcas de agua robustas en imágenes digitales con formato *Escuela Superior de Ingeniería Mecánica y Eléctrica ESIME Zacatenco, IPN*. 2005.
[26] E. Brannock and et al. Watermarking with Wavelets: Simplicity leads to robustness *Proceedings of IEEE Southeastcon*, pages 587–592. 2008.
[27] S. Lee and et al. Reversible Image Watermarking Based on Integer-to-Integer Wavelet Transform *IEEE Transactions on Information Forensics and Security, Vol. 2, No. 3*, pages 321–330. 2007.
[28] S. P. Mayit and M. K. Kundu. Performance improvement in spread spectrum image watermarking using Wavelets *International Journal of Wavelets, Multiresolution and Information Processing, Vol. 9*, pages 1-33. 2011.
[29] M. K. Hu Visual Pattern Recognition by Moment Invariants *IRE Transactions on Information Theory, Vol. 8, No. 2*, pages 179–187. 1962.
[30] G. Schouten Thesis: Luminance-Brightness Mapping: the Missing Decades *Technische Universiteit Eindhoven*. 1992.
[31] A. B. Watson. DCT quantization matrices visually optimized for individual images *Proceedings of the SPIE Human Vision, Visual Processing, and Digital Display IV*. 1993.
[32] B. Escalante, P. López, and J. L. Silvan SAR Image Classification with a Directional-Oriented Discrete Hermite Transform *Proceedings SPIE Image and Signal Processing for Remote Sensing VII*. 2002.
[33] J. R. Ohm Bildsignalverarbeitung fuer Multimadia-Systeme `http://bs.hhi.de/user/ohm/download/bvm-kap1&2.pdf`
[34] B. L. Gunjal and R. R. Manthalkar Discrete Wavelet Transform based Strongly Robust Watermarking Scheme for Information Hiding in Digital Images *Third International Conference on Emerging Trends in Engineering and Technology*, pages 124–129. 2010.

Sandra L. Gomez Coronel Received her BS degree in Communications and Electronics Engineering from Escuela Superior de Ingeniería Mecánica y Eléctrica at Instituto Politécnico Nacional (IPN) in 2003, her MSc degree in Telecommunications Engineering in 2008 from SEPI ESIME Zacatenco (IPN) and her PhD degree from Universidad Nacional Autónoma de México in 2014. Her research interests include digital signal processing, specifically image processing and their applications. Also she has several years as bachelor professor at UPIITA (IPN).

Marco A. Acevedo Mosqueda He was born in Mexico City in July 19th, 1968. He received his BS degree in Communications and Electronics Engineering in 1992 and his MSc degree with specialization in Electronics in 1996 from Escuela Superior de Ingeniería Mecánica y Eléctrica at Instituto Politécnico Nacional. Currently, he is a professor in ESIME. His main research area is Digital Signal Processing and Telecommunications.

Ma. Elena Aevedo Mosqueda Received her BS degree in Engineering with specialization in Computing from the Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME) at Instituto Politécnico Nacional (IPN) in 1996. She has been teaching at ESIME since 1994. She received her MSc degree with specialization in Computing from the Centro de Investigación y de Estudios Avanzados (CINVESTAV) in 2001. She received her PhD from the Centro de Investigación en Computación (CIC) at IPN in 2006. Her main research area is Artificial Intelligence and Associative Memories.

Ricardo Carreño Aguilera He studied a bachelor in Communications and Electronics Engineering with specialization in control at ESIME-ZACATENCO IPN, a master in business administration at the University of Querétaro and a PhD in systems engineering in the post-grade and research area at ESIME-ZACATENCO IPN. Currently he is developing research in complex systems and artificial intelligence.

ISSN 2395-8618

Comparative Study of Computational Tools for Hub Gene Selection from Genetic Network using Microarray Data

Bijeta Mandal, Saswati Mahapatra, and Tripti Swarnkar

***Abstract*—Selection of genes associated to complex diseases has been a challenging task in the field of bioinformatics. Through various studies it has been concluded that selection of highly connected intramodular hub genes in a co-expression network analysis approach leads to more biologically relevant gene lists. In this paper, we have assess the empirical performance of three existing network reconstruction methods Weighted Gene Correlation Network Analysis (WGCNA), Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE), Graphical Lasso (GLASSO). The study compares the extracted hub genes from estimated networks on the prostate cancer dataset based on two criteria: the first criterion evaluates the biological enrichment and the second criterion evaluates the statistical validation, prediction accuracy. The result suggests, though there is considerable amount of heterogeneity, randomness and variability in structures of networks estimated using different reconstruction methods, our findings provides evidence for similarity in hub genes selection. These findings after network analysis can provide an intuitive insight into selection of network estimation methods for specific range of gene expression in microarray datasets. *Index Terms*—Gene Selection, Intramodular hub gene, Co-expression network, Genetic Network, Network reconstruction, Network analysis, Microarray**

I. Introduction

Understanding the relationship among genes, is extremely fundamental with a specific end goal to analyze genomic data. Gene expression data can be productively dissected with network methods characterizing clusters of interconnected genes [1], with edges capturing interactions at different levels. Genetic interactions hypothesizes activities of biological pathway, cellular response [2], acknowledging elements of genes from their reliance on different genes [3], distinguishing novel biomarkers [4] and more precise classification methods [5]. The degree of interactions in the clusters are significantly higher than an irregular network exhibiting indistinguishable degree distribution [6]. Diverse statistical and bioinformatics techniques can be applied directly to microarray data to estimate networks of genetic interactions in different cellular states or disease stages with a common motive to glean an edge among a pair of genes by considering a cue of association, which is pivotal in different network reconstruction method [7]. Associations in network can be classified into two: Marginal associations that ignores the nearness of other genes while estimating an edge between genes and conditional associations that considers impact of nearness of other genes while concluding an edge between genes. Focusing on intramodular hubs instead of whole network hubs for co-expression network applications leads to better results of clinical significance [8] a key factors in a network architecture [9], and are often strongly enriched in specific functional categories or cell markers [10]. Empirical evidence shows gene selection based on intramodular connectivity leads to biologically more informative gene lists focusing on the relationship between modules and the sample trait [1], [11], [12], that prompts gene connectivity can be used for identifying hubs and differentially connected genes [11], [13], [14] for finding biological information embedded in microarray data [13], [14]. Our comparative study includes a comparison of three computational methods with publicly available software, Weighted Gene Correlation Network Analysis (WGCNA) [15], Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) [16], Graphical Lasso (GLASSO) [17] for reconstruction of genetic networks with undirected edges as it is not possible to estimate directed edges with observational data alone [18]. Different computational tools are implemented to a benchmark dataset to analyze similarities and differences in estimated networks and their performances in terms of intramodular hub genes. Finally, the presence of cancer related genes and their influence in specific cancer type using NCBI database and DAVID [19] has been studied. The result provides a insight into the presence of cancer-related genes in the hub gene modules found in known biological networks [20] and also helps in selection of most efficient network estimation method. The rest of the paper includes detail of methods for network reconstruction, proposed model, results analysis and discussion of our findings for reconstructed genetic networks, and future research scope.

Manuscript received on September 12, 2016, accepted for publication on December 17, 2016, published on June 30, 2017.

The authors are with the Department of Computer Application, Siksha 'O'Anusandhan University, Bhubaneswar 751 030, India (e-mail: bijetamandal@gmail.com, saswatimohapatra@soauniversity.ac.in, triptiswarnakar@soauniversity.ac.in).

II. METHODS AND MATERIALS

A. Methods

WGCNA [15] is a genetic network reconstruction tool based on marginal measure of correlation patterns among

ISSN 2395-8618

genes that incorporates functions for finding modules of highly correlated genes. In WGCNA, gene significance and module eigengene or intramodular hub gene based connectivity among genes facilitate gene screening methods to identify candidate biomarkers, and can be used to generate testable hypotheses for validation in independent datasets [21]. WGCNA is implemented using R software. ARACNE [16] is based on removal of non-linear similarities among expression levels for a pair of genes. The algorithm computes pair wise mutual information MI_{ij} for each pair of genes *i* and *j*, and applies DPI (Data Processing Inequality) as a pruning step for removal of the false positive edges corresponding to indirect interactions in the network. ARACNE is classified as a method based on a blend of marginal and conditional associations and is implemented using package minet [22] in the Bioconductor. Graphical lasso (GLASSO) [17] is a network estimation tool based on sparsity inducing penalties i.e. lasso penalty assuming multivariate normality for random variables [7]. GLASSO estimates inverse covariance matrix by maximizing the l_1- penalized log likelihood function to construct a sparse graphs of conditional independence relations among the genes. The tuning parameter ρ is a positive number controlling the degree of sparsity. It is implemented with package glasso in R.

B. Datasets

The prostate cancer microarray dataset for homo sapiens consisting 104 samples and 20000 genes with 6 variants in samples [20], is utilized in our study to show the effectiveness of our proposed model, obtained from NCBIs Gene Expression Omnibus (GEO). For the purpose of statistical analysis the samples are categorized into two, 70 diseased samples and 34 normal samples. Biological dataset adopted for validation collected from the NCBI gene database includes 7238 cancer-related genes and 2202 prostate cancer genes.

C. Proposed Model

Block diagram in Fig 1, represents the schematic work flow of the proposed hub genes selection model.

1) Data Preprocessing.: The methods are being implemented in R software using different R packages. The preprocessing of data includes cleaning of data by removal of genes with large number of missing values. Hierarchical clustering is performed for finding sample outliers in the samples. Missing values of a gene are replaced with the mean value of observed data. The genes are filtered based on their variances across diseased and normal samples producing 100 samples and 14689 probes. Due to technical limitations regarding memory allocation during GLASSO implementation (System specification:12 GB RAM) we had to confined the number of probes not more than 10000 for different computational tool implementation.

Computational Tools Implementation.:

a) WGCNA.: Pearson correlation S_{ij} is calculated for the gene expression profile and are then transformed into adjacency matrix by applying a power adjacency function $|S_{ij}|\,\beta$, where the exponent β is the power estimate to obtain a scale-free topology [23]. Further co-expression values are converted to the topology overlap measure (TOM), that facilitates the identification of gene modules. The output of the implementation showed 19 modules. Based on high module membership and intra modular connectivity hub gene modules are selected.

b) ARACNE.: Mutual information (MI) is evaluated between each pair of genes and is taken as input to the aracne() function for network estimation. The number of the edges are controlled by thresholding the value of MI for each pair of genes in the network. The output of the implementation showed 1 module. For analysis and comparison with network estimation from other tools, connectivity of each node is considered.

c) GLASSO.: Covariance matrix is calculated between each pair of genes and taken as input to glasso() function to calculate an inverse covariance matrix for network estimation. In our implementation we have opted for two variations of GLASSO i.e. defining diagonal of inverse covariance to be penalized or not. The output of the implementation showed 1 module each for both the variations of GLASSO implementation considering the module constraint of minimum 25 genes, taken as standard in WGCNA.

2) Extraction of Hub Gene Modules.: A hub gene module with high intramodular connectivity can be considered as a gene module with strongly interacting genes. Study shows genes with higher module membership show higher intramodular connectivity and are more biologically significant [15]. A set of twenty top ranked genes are extracted from each module to create hub gene modules for further analysis. Integrated modules shows improved classification performance in gene selection [20], so we have selected five top ranked genes from individual hub gene module to construct a integrated hub module.

3) Performance evaluation of selected hub genes.:

a) Statistical Analysis.: Predictive accuracy of the hub genes are measured in terms of Matthews coefficient correlation (MCC), as it is a measure of quality of binary classification. [24], [25]. MCC, overall accuracy, sensitivity, specificity, precision and f-measure are adopted for statistical analysis in comparison to the known true classes [25].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (1)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

ISSN 2395-8618

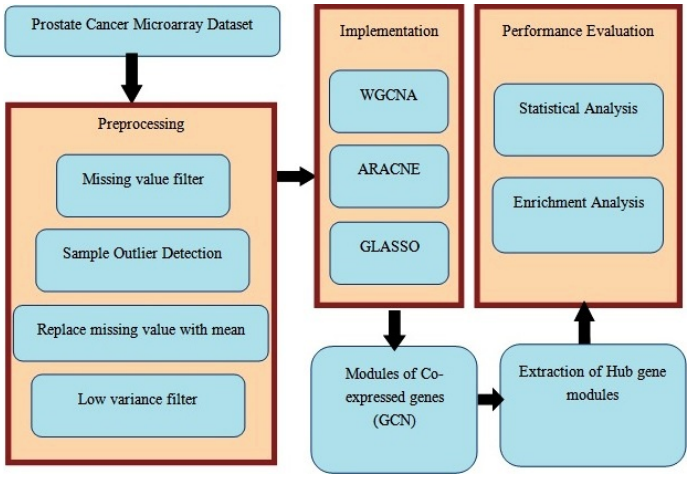

Fig. 1. Steps for gene selection in proposed model

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$F - measure = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{6}$$

where TP is number of true-positive samples, TN is count for true-negative samples, FP is number of false-positive samples and FN is number of false-negative samples.

b) Enrichment Analysis.: The biological significance of the selected hub genes are firstly validated with the percentage of disease-related genes in them and secondly the results are validated by summarizing the genes belonging to an enriched functional category measured in terms of p-value [26] and fold enrichment, of enriched attributes (EA) using DAVID [19].

III. RESULT AND DISCUSSION

In the paper we have performed three step evaluation of the selected modules for the hub genes selection.
(i) Comparison of modules based on the graph density of hub gene modules.
(ii) Effectiveness of selected hub gene modules are analyzed in terms of prediction accuracy.
(iii) Biological significance is analyzed involving presence of disease related genes and enriched attributes.
After applying the progression (ii) of our proposed model for GCN construction utilizing distinctive computational strategies, brought about 19 modules in WGCNA ,their genes are ranked predicated on their intra modular connectivity. An arrangement of twenty top positioned genes are extracted from every module to extracted 19 hub gene modules for further analysis, that tallies to cull of 439 genes. Assuming integrated modules shows amended relegation performance in gene glean [20], We have sorted out five top ranked genes from individual hub gene module to contrive a integrated hub module Hub5 with 125 genes. Hub genes and subset of hub gene modules are constructed from the modules estimated using ARACNE(A1-603 genes, A2-179 genes) and GLASSO(for penalized diagonal false: F1-497 genes, F2-134 genes, for penalized diagonal true: T1-497 genes, T2-140 genes implementation, by considering the heterogeneity in degree distribution for network estimates utilizing distinctive computational tools and number of genes selected as hub genes in WGCNA for individual and integrated modules as standard. The distinct co-expressed gene modules and integrated modules constructed using distinctive computational strategies of our approach are designated as following: Bl Black, B Blue, Br Brown, C Cyan, G Green, GY Green Yellow, G60 Grey 60, LC Light Cyan, LG Light Green, LY Light Yellow, M Magenta, MB Midnight Blue, P Pink, Pu Purple, R Red, S Salmon, Tn Tan, T Turquoise, Y Yellow, Hub5 are the co-expressed hub gene modules obtained using WGCNA based network construction approach. A1, A2 are the co-expressed hub gene modules obtained using ARACNE based network construction approach and F1, F2, T1, T2 are the co-expressed hub gene modules obtained using GLASSO based network construction approach.

A. Graph Density Analysis

We surmise that the more precise and dense the gene module, the higher the quality measure [20]. In Fig. 2 and Fig. 3, we have summarized the results for all hub gene modules from different computational tools in terms of graph density (the ratio between number of edges and number of nodes/genes) for prostate cancer dataset. After obtaining graph density measure for different co-expressed hub gene module from different computational tools, we filtered 6 different hub gene modules in WGCNA (five individual modules and one integrated hub gene module) for prostate dataset. As number of modules in ARACNE and GLASSO implementation is very less so all the hub gene modules are considered for the study. The selected hub gene modules show comparatively high graph density with respect to the intramodular connectivity. Thus, these selected hub gene modules, are further considered for statistical and biological analysis.

ISSN 2395-8618

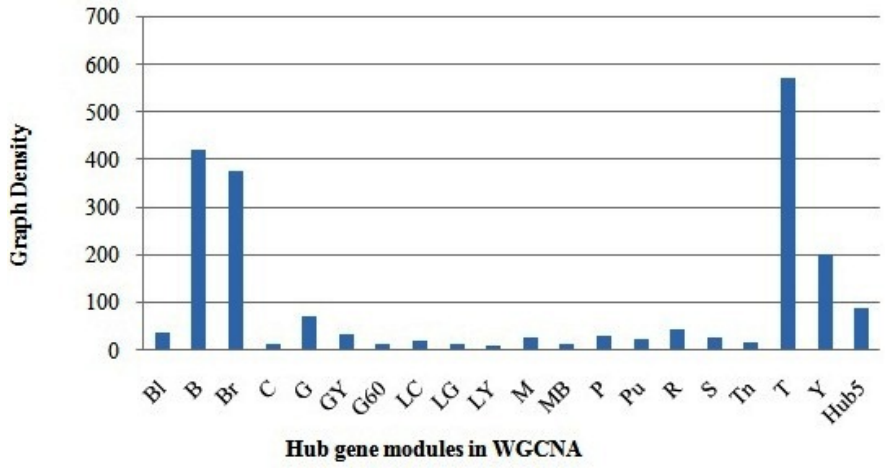

Fig. 2. Graph density of Hub gene module for WGCNA

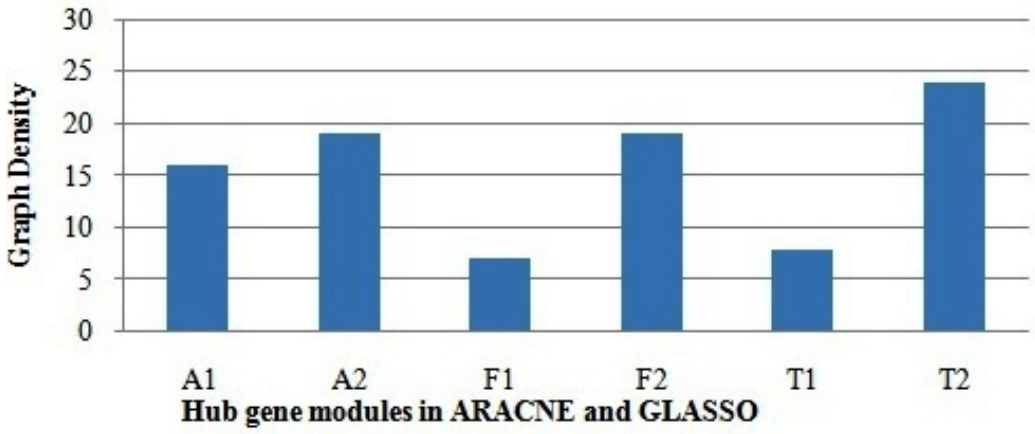

Fig. 3. Graph density of Hub gene module for ARACNE and GLASSO

B. Classification performance

Hub genes are high degree nodes that incline to play a consequential role in the functional modules [27]. The performance of the hub gene modules is evaluated in terms of predictive accuracy as listed in Table 1 for prostate cancer dataset . The kNN (k=3), Random Forest and SVM with tenfold cross validation are applied as classifiers [20]. From Table 1, we observed few hub gene modules in WGCNA (Blue,Hub5) shows better results than of the individual hub gene modules, for ARACNE (A1,A2) and for GLASSO (F1, F2, T1, F1) the results are good.

C. Biological Significance analysis

The biological analysis of co-expressed hub gene modules are based on the fol-lowing criteria:

1) Disease-related genes analysis.: Fig 4, illustrates the efficacy of hub gene selection in terms of identifying disease-related genes represented as the percentage of studied cancer (prostate) related genes in each significant hub gene module. It is been observed that the hub gene modules with enhanced prediction accuracy have high fraction of co-expressed cancer-related genes. They are being considered as significant for further study for genes mostly related with disease.

2) Analysis of enriched attributes associated with prostate cancer hub gene module.: The biological significance is evaluated in terms of percentage of genes related with specific relevant biological process in each hub gene module and are shown in Table 2 for prostate cancer data. The biological significance of the genes belonging to an enriched functional category can be measured in terms of p-value [26]. The results are validated using p-value value cut-off of 5×10^2 and fold enrichment (FE) 1.5 [6], of enriched attributes/functions (EA) in our study. Since DAVID gene ID is unique per gene, it is more accurate to use DAVID ID to present the gene-annotation association by removing any redundancy in user gene list. Interestingly, Hub5, A1, F1, F2, T1, T2 shows relatively large number of EAs satisfying the p-value and FE cut-off.

ISSN 2395-8618

TABLE I
BIOLOGICAL FUNCTIONAL ANALYSIS OF GENES IN HUB GENE MODULES IN TERMS OF ENRICHED ATTRIBUTE COUNT

M	NG	CL	3NN						RF						SVM					
			Sen	Spec	Prec	Fm	Mcc	Acc	Sen	Spec	Prec	Fm	Mcc	Acc	Sen	Spec	Prec	Fm	Mcc	Acc
B	20	N	0.62	0.86	0.07	0.66	S0.50	0.78	0.56	0.97	0.91	0.69	0.62	0.83	0.35	1.00	1.00	0.52	0.51	0.78
		P	0.86	0.62	0.81	0.84	0.50		0.97	0.56	0.81	0.88	0.62		1.00	0.35	0.75	0.86	0.51	
Br	21	N	0.77	0.77	0.63	0.69	0.52	0.77	0.65	0.85	0.69	0.67	0.50	0.78	0.12	0.88	0.33	0.17	-0.01	0.62
		P	0.77	0.77	0.86	0.82	0.52		0.85	0.65	0.82	0.84	0.50		0.88	0.12	0.66	0.75	-0.01	
G	24	N	0.65	0.79	0.61	0.63	0.43	0.74	0.41	0.88	0.64	0.50	0.33	0.72	0.38	0.94	0.77	0.51	0.41	0.75
		P	0.79	0.65	0.81	0.80	0.43		0.88	0.41	0.74	0.81	0.33		0.94	0.38	0.75	0.51	0.41	
T	21	N	0.53	0.74	0.51	0.52	0.27	0.67	0.47	0.86	0.64	0.54	0.37	0.73	0.00	1.00	0.00	0.00	0.00	0.66
		P	0.74	0.53	0.75	0.75	0.27		0.86	0.47	0.76	0.81	0.37		1.00	0.00	0.66	0.80	0.00	
Y	22	N	0.47	0.79	0.53	0.50	0.27	0.68	0.35	0.94	0.75	0.48	0.38	0.74	0.03	0.99	0.50	0.06	0.05	0.66
		P	0.79	0.47	0.74	0.77	0.27		0.94	0.35	0.74	0.83	0.38		0.99	0.03	0.66	0.79	0.05	
Hub5	125	N	0.79	0.96	0.90	0.84	0.77	0.90	0.68	0.96	0.89	0.77	0.68	0.86	0.71	0.94	0.86	0.77	0.68	0.86
		P	0.96	0.79	0.90	0.93	0.77		0.96	0.68	0.85	0.90	0.68		0.94	0.71	0.86	0.90	0.68	
A1	603	N	0.91	0.86	0.78	0.84	0.75	0.88	0.68	0.99	0.96	0.79	0.73	0.88	0.59	0.97	0.91	0.71	0.64	0.84
		P	0.86	0.91	0.95	0.91	0.75		0.99	0.68	0.86	0.92	0.73		0.97	0.59	0.82	0.89	0.64	
A2	179	N	0.85	0.85	0.74	0.80	0.68	0.85	0.62	0.94	0.84	0.71	0.61	0.83	0.53	0.97	0.90	0.67	0.59	0.82
		P	0.85	0.85	0.92	0.88	0.68		0.94	0.62	0.83	0.88	0.61		0.97	0.53	0.80	0.88	0.59	
F1	497	N	0.91	0.92	0.86	0.89	0.83	0.92	0.74	0.97	0.93	0.82	0.75	0.89	0.71	0.97	0.92	0.80	0.73	0.88
		P	0.92	0.91	0.95	0.94	0.83		0.97	0.74	0.88	0.92	0.75		0.71	0.71	0.87	0.91	0.73	
F2	134	N	0.91	0.91	0.84	0.87	0.81	0.91	0.79	0.97	0.93	0.86	0.80	0.91	0.82	0.96	0.90	0.86	0.80	0.91
		P	0.91	0.91	0.95	0.93	0.81		0.97	0.79	0.90	0.93	0.80		0.96	0.82	0.91	0.93	0.80	
T1	497	N	0.91	0.92	0.86	0.89	0.83	0.92	0.74	0.97	0.93	0.82	0.75	0.89	0.71	0.97	0.92	0.80	0.73	0.88
		P	0.92	0.91	0.95	0.94	0.83		0.97	0.74	0.88	0.92	0.75		0.97	0.71	0.87	0.91	0.73	
T2	140	N	0.91	0.91	0.84	0.87	0.81	0.91	0.82	0.97	0.93	0.88	0.82	0.92	0.79	0.97	0.93	0.86	0.80	0.91
		P	0.91	0.91	0.95	0.93	0.81		0.97	0.82	0.91	0.94	0.82		0.97	0.79	0.90	0.93	0.80	

Bold hub gene module specifies the hub gene modules showing comparable good predictive accuracy measures. NG number of genes, Cl Class label, 3NN 3 nearest neighbors, RF random forest, SVM support vector machine, Sen sensitivity, Spec specificity, Prec precision, Fm F-measure, Mcc Matthews correlation coefficient; Acc prediction accuracy, N negative (normal) sample, P positive (prostate cancer) sample

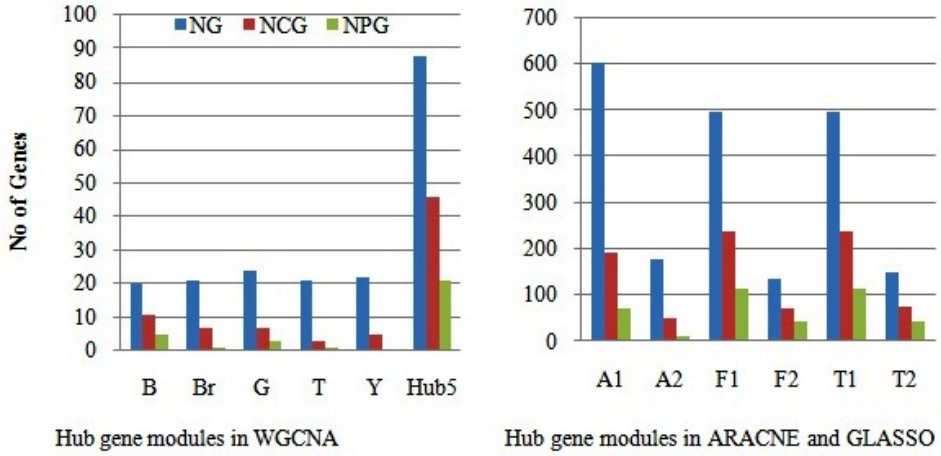

Fig. 4. Biological significance study of the hub gene modules in terms of the presence of disease-related genes for prostate cancer dataset in WGCNA, ARACNE and GLASSO. NG number of genes in hub gene module, NCG number of Cancer genes in hub gene module, NPG number of Prostate Cancer genes in hub gene module

ISSN 2395-8618

Few of the biological functions more related to the disease are also found enriched in the modules. These processes mainly include transcription, translation and RNA binding that plays an important role in protein regulation. Acetylation and phosphoproteins are known to play a vital role in genetics modification that occurs in cancer.The dysregulation of cell cycle, spliceosome and focal adhesion plays pivotal role in cancer metastasis. Regulation of apoptosis, UBL conjugation are important parts of programmed cell death and have significant change in cancer progression [20].

IV. CONCLUSION

The advantage of focusing on intramodular hub genes instead of whole network of co-expressed genes leads to better selection of biologically enriched and statically significant biomarkers. The study shows the comparison of gene selection using three widely used standard computational tools. We have evaluated the selected hub gene modules for three different benchmark methods based on their graph density, predic-tion accuracy and presence of enriched attributes. Considering graph density as meas-ure, modules formed in WGCNA are more dense than modules estimated from other tools. The statistical analysis of selected modules based on graph density shows, modules in ARACNE, GLASSO and integrated module in WGCNA have compara-tively similar class performance and outperforming the individual modules in WGCNA showing moderate accuracy. The modules in GLASSO are biologically more significant with respect to presence of enriched attributes than the modules in ARACNE and WGCNA. All the standard computational methods used in the study are showing similar performance, at the same time GLASSO and ARACNE are show-ing more computational complexity based on size of the modules created. The hub gene selected using different computational tools may further be provided to different known networks which may provide greater insights into the fundamental biology and pathogenesis of the disease.

REFERENCES

[1] P. Langfelder, P. S. Mischel, and S. Horvath, "When is hub gene selection better than standard meta-analysis?" *PloS one*, vol. 8, no. 4, p. e61505, 2013.

[2] A. B. Parsons, R. L. Brost, H. Ding, Z. Li, C. Zhang, B. Sheikh, G. W. Brown, P. M. Kane, T. R. Hughes, and C. Boone, "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways," *Nature biotechnology*, vol. 22, no. 1, pp. 62–69, 2004.

[3] P. Ye, B. D. Peyser, X. Pan, J. D. Boeke, F. A. Spencer, and J. S. Bader, "Gene function prediction from congruent synthetic lethal interactions in yeast," *Molecular systems biology*, vol. 1, no. 1, 2005.

[4] L. Chen, J. Xuan, R. B. Riggins, R. Clarke, and Y. Wang, "Identifying cancer biomarkers by network-constrained support vector machines," *BMC systems biology*, vol. 5, no. 1, p. 1, 2011.

[5] P. Hu, S. B. Bull, and H. Jiang, "Gene network modular-based classification of microarray samples," *BMC bioinformatics*, vol. 13, no. 10, p. 1, 2012.

[6] T. Swarnkar, S. N. Simoes, D. C. Martins, A. Anurak, H. Brentani, R. F. Hashimoto, and P. Mitra, "Multiview clustering on ppi network for gene selection and enrichment from microarray data," in *Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on*. IEEE.

[7] N. Sedaghat, T. Saegusa, T. Randolph, and A. Shojaie, "Comparative study of computational methods for reconstructing genetic networks of cancer-related pathways," *Cancer Inform*, vol. 13, no. Suppl 2, 2014.

[8] S. Horvath, B. Zhang, M. Carlson, K. Lu, S. Zhu, R. Felciano, M. Laurance, W. Zhao, S. Qi, Z. Chen *et al.*, "Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target," *Proceedings of the National Academy of Sciences*, vol. 103, no. 46, pp. 17 402–17 407, 2006.

[9] E. Almaas, "Biological impacts and context of network theory," *Journal of Experimental Biology*, vol. 210, no. 9, pp. 1548–1558, 2007.

[10] M. C. Oldham, G. Konopka, K. Iwamoto, P. Langfelder, T. Kato, S. Horvath, and D. H. Geschwind, "Functional organization of the transcriptome in human brain," *Nature neuroscience*, vol. 11, no. 11, pp. 1271–1282, 2008.

[11] T. F. Fuller, A. Ghazalpour, J. E. Aten, T. A. Drake, A. J. Lusis, and S. Horvath, "Weighted gene coexpression network analysis strategies applied to mouse weight," *Mammalian Genome*, vol. 18, no. 6-7, pp. 463–472, 2007.

[12] P. Langfelder, L. W. Castellani, Z. Zhou, E. Paul, R. Davis, E. E. Schadt, A. J. Lusis, S. Horvath, and M. Mehrabian, "A systems genetic analysis of high density lipoprotein metabolism and network preservation across mouse models," *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, vol. 1821, no. 3, pp. 435–447, 2012.

[13] X. Xu and A. Zhang, "Selecting informative genes from microarray dataset by incorporating gene ontology," in *Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)*. IEEE, 2005, pp. 241–245.

[14] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 1, 2006.

[15] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 1, 2008.

[16] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006.

[17] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[18] A. Shojaie, A. Jauhiainen, M. Kallitsis, and G. Michailidis, "Inferring regulatory networks by combining perturbation screens and steady state gene expression profiles," *PloS one*, vol. 9, no. 2, p. e82393, 2014.

[19] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2009.

[20] T. Swarnkar, S. N. Simões, A. Anura, H. Brentani, J. Chatterjee, R. F. Hashimoto, D. C. Martins, and P. Mitra, "Identifying dense subgraphs in protein–protein interaction network for gene selection from microarray data," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 4, no. 1, pp. 1–18, 2015.

[21] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS comput biol*, vol. 4, no. 8, p. e1000117, 2008.

[22] P. E. Meyer, F. Lafitte, and G. Bontempi, "minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information," *BMC bioinformatics*, vol. 9, no. 1, p. 1, 2008.

[23] A. Li and S. Horvath, "Network neighborhood analysis with the multi-node topological overlap measure," *Bioinformatics*, vol. 23, no. 2, pp. 222–231, 2007.

[24] P. Dao, K. Wang, C. Collins, M. Ester, A. Lapuk, and S. C. Sahinalp, "Optimally discriminative subnetwork markers predict response to chemotherapy," *Bioinformatics*, vol. 27, no. 13, pp. i205–i213, 2011.

[25] J. Ahn, Y. Yoon, C. Park, E. Shin, and S. Park, "Integrative gene network construction for predicting a set of complementary prostate cancer genes," *Bioinformatics*, vol. 27, no. 13, pp. 1846–1853, 2011.

[26] A. Ghosh, B. C. Dhara, and R. K. De, "Selection of genes mediating certain cancers, using a neuro-fuzzy approach," *Neurocomputing*, vol. 133, pp. 122–140, 2014.

[27] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

ISSN 2395-8618

TABLE II
BIOLOGICAL FUNCTIONAL ANALYSIS OF GENES IN HUB GENE MODULES IN TERMS OF ENRICHED ATTRIBUTE COUNT

CM	Modules	NG	DC	No of EA
WGCNA	B	20	15	14
	Br	21	18	1
	G	24	19	10
	T	21	13	1
	Y	22	9	1
	Hub5	125	88	120
ARACNE	A1	603	421	219
ARACNE	A2	179	126	53
GLASSO	F1	497	367	580
GLASSO	F2	134	103	369
GLASSO	T1	497	367	580
GLASSO	T2	140	108	340

Computational method, NG number of genes, DC DAVID ID count EA enriched attribute

ISSN 2395-8618

Journal Information and Instructions for Authors

I. Journal Information

Polibits is a half-yearly open-access research journal published since 1989 by the *Centro de Innovación y Desarrollo Tecnológico en Cómputo* (CIDETEC: Center of Innovation and Technological Development in Computing) of the *Instituto Politécnico Nacional* (IPN: National Polytechnic Institute), Mexico City, Mexico.

The journal has double-blind review procedure. It publishes papers in English and Spanish (with abstract in English). Publication has no cost for the authors.

A. Main Topics of Interest

The journal publishes research papers in all areas of computer science and computer engineering, with emphasis on applied research. The main topics of interest include, but are not limited to, the following:

- Artificial Intelligence
- Natural Language Processing
- Fuzzy Logic
- Computer Vision
- Multiagent Systems
- Bioinformatics
- Neural Networks
- Evolutionary Algorithms
- Knowledge Representation
- Expert Systems
- Intelligent Interfaces
- Multimedia and Virtual Reality
- Machine Learning
- Pattern Recognition
- Intelligent Tutoring Systems
- Semantic Web
- Robotics
- Geo-processing
- Database Systems
- Data Mining
- Software Engineering
- Web Design
- Compilers
- Formal Languages
- Operating Systems
- Distributed Systems
- Parallelism
- Real Time Systems
- Algorithm Theory
- Scientific Computing
- High-Performance Computing
- Networks and Connectivity
- Cryptography
- Informatics Security
- Digital Systems Design
- Digital Signal Processing
- Control Systems
- Virtual Instrumentation
- Computer Architectures

B. Indexing

The journal is listed in the list of excellence of the CONACYT (Mexican Ministry of Science) and indexed in the following international indices: Web of Science (via SciELO citation index), LatIndex, SciELO, Redalyc, Periódica, e-revistas, and Cabell's Directories.

There are currently only two Mexican computer science journals recognized by the CONACYT in its list of excellence, *Polibits* being one of them.

II. Instructions for Authors

A. Submission

Papers ready for peer review are received through the Web submission system on www.easychair.org/conferences/?conf=polibits1; see also updated information on the web page of the journal, www.cidetec.ipn.mx/polibits.

The papers can be written in English or Spanish. In case of Spanish, author names, abstract, and keywords must be provided in both Spanish and English; in recent issues of the journal you can find examples of how they are formatted.

The papers should be structures in a way traditional for scientific paper. Only full papers are reviewed; abstracts are not considered as submissions. The review procedure is double-blind. Therefore, papers should be submitted without names and affiliations of the authors and without any other data that reveal the authors' identity.

For review, a PDF file is to be submitted. In case of acceptance, the authors will need to upload the source code of the paper, either Microsoft Word or LaTeX with all supplementary files necessary for compilation. Upon acceptance notification, the authors receive further instructions on uploading the camera-ready source files.

Papers can be submitted at any moment; if accepted, the paper will be scheduled for inclusion in one of forthcoming issues, according to availability and the size of backlog.

See more detailed information at the website of the journal.

B. Format

The papers should be submitted in the format of the IEEE Transactions 8x11 2-column format, see http://www.ieee.org/publications_standards/publications/authors/author_templates.html. (while the journal uses this format for submissions, it is in no way affiliated with, or endorsed by, IEEE). The actual publication format differs from the one mentioned above; the papers will be adjusted by the editorial team.

There is no specific page limit: we welcome both short and long papers, provided that the quality and novelty of the paper adequately justifies its length. Usually the papers are between 10 and 20 pages; much shorter papers often do not offer sufficient detail to justify publication.

The editors keep the right to copyedit or modify the format and style of the final version of the paper if necessary.

See more detailed information at the website of the journal.