Are my Children Old Enough to Read these Books? Age Suitability Analysis

Franz Wanner, Johannes Fuchs, Daniela Oelke, and Daniel A. Keim

***Abstract*—In general, books are not appropriate for all ages, so the aim of this work was to find an effective method of representing the age suitability of textual documents, making use of automatic analysis and visualization. Interviews with experts identified possible aspects of a text (such as 'is it hard to read?') and a set of features were devised (such as linguistic complexity, story complexity, genre) which combine to characterize these age related aspects. In order to measure these properties, we map a set of text features onto each one. An evaluation of the measures, using Amazon Mechanical Turk, showed promising results. Finally, the set features are visualized in our age-suitability tool, which gives the user the possibility to explore the results, supporting transparency and traceability as well as the opportunity to deal with the limitations of automatic methods and computability issues.**

***Index Terms*—Information interfaces and presentation, information search and retrieval.**

I. Introduction

TWITTER messages, blog posts, customer reviews, and other user-generated content in the internet provide a wealth of information for companies and potential customers to learn about the strengths and weaknesses of different products. Studies have shown that about 81% of the Internet users in the U.S. have done online research on a product at least once [1]. In the last years, many text analysis approaches were developed that support the user in mining these resources. Automatic algorithms for opinion and sentiment detection permit to process a set of customer reviews automatically and present a summary of the product's most liked or disliked features.

This approach works well for many types of products. However, there are purchase decisions that are not adequately supported by the available methods. For example, before buying a book many potential readers would like to see if the writing style suits their taste. Some online stores meet this need by offering a "Look Inside" functionality that allows you to read some pages of the book. But this often is not enough to determine what age a book is suitable for. To assess this more than just the writing style needs to be taken into account.

For many books, the retail market and sometimes also the publishers provide a recommendation for the reader's age. However, often this is arguable. For example, the whole series of "Harry Potter" is recommended as being suitable for readers at the age of 9 to 12. Critics remarked that there is clear increase in violence and blood-curdling fragments in the later books of the series. Furthermore, the length of the book changed from 300 pages in the first volume to more than 780 in the final book of the series. It was therefore encouraged to rethink whether the books should really be all recommended for the same age range. Our interviews in german book stores confirmed this impression: at least some retailers shared this subjective view about the book.

Asked what aspects should be taken into account when determining the age group that a book is suitable for, the interviewed retailers suggested to take a look at the following parameters: (a) The difficulty of the writing style, (b) the complexity of the story, (c) the topics that are covered, (d) the emotions that are aroused, and finally (e) the ratio between pictures and textual content as well as other physical aspects such as the font size that is used.

In this paper, we present an approach that computationally assesses these five aspects. Rating books with an automatic algorithm comes with the advantage that it is independent of economic interests and individual opinions and positions. By measuring the different aspects separately and subsequently visualizing the result, it becomes possible to weight the different influences as desired. This permits to take individual preferences and special needs of the reader into account.

The paper is structured as follows: After a discussion of related work in section II, we introduce the different features for measuring age suitability in section III. With the help of the Amazon Mechanical Turk [2], a ground-truth data set was established that is then used in section IV to evaluate the features. Finally, a multi-view dashboard visualization is provided that allows the user to explore the detailed information that was extracted about the book (section V). Section VI concludes the paper.

II. Related Work

A. Related Work for Features Approximating Age Suitability

Subjectivity analysis is the recognition of opinion-oriented language in order to distinguish it from objective language. Sub-areas of subjectivity analysis are opinion or sentiment analysis. Many approaches and definitions can be found in [3]. However, the detection of emotion is slightly different. Important here is the determination of the expressed emotion. In [4] and [5] this was done for web news. The work in the

Manuscript received October 27, 2010. Manuscript accepted for publication January 28, 2011.

The authors are with the University of Konstanz, 78457 Konstanz, Germany (e-mail: wanner@dbvis.inf.uni-konstanz.de, Johannes.Fuchs@uni-konstanz.de, oelke@dbvis.inf.uni-konstanz.de, keim@dbvis.inf.uni-konstanz.de).

area of topic detection is tremendous and the focus lies on methods to detect and track events automatically. However, our goal is to get the specific topic of a book. Nallapati [6] compared the content of news articles by means of four categories. When the categories overlap sufficiently, then the compared documents build a topic. Another approaches are more appropriate for our needs determing topics in advance. The text classification algorithms of Green [7], Scott [8] or Hotho et al. [9] use WordNet, a lexical database. The advantage of such an approach is to provide semantical knowledge to the classification algorithm. Further methods and techniques can be found in the book of James Allan [10]. Text properties can be special in the sense that they do not measure a property that is in the text, but rather an "effect" that is caused by the text [11]. The story complexity can be seen as an effect, caused by many different characters and a fragmented story. Beside the already introduced readability of Oelke et al.,there are different algorithms to determine the readability of textual documents. Popular ones amongst others are the Gunning Fog [12] or the Flesch-Kincaid Readability Test [13]. It is common to all these measures that they base on statistical characteristics of the analyzed text. Additionally, we measure the vocabulary richness. This has been mainly used in the area of authorship attribution, for example [14] or [15].

B. Visual Approaches for Document Analysis

Full automatic algorithms hit their limit when human knowledge is required and in order to understand a document, knowledge of the world and human interpretation is needed [16]. This is the point where *Visual Analytics* can help. The aim of Visual Analytics is to make the way of processing data and information transparent for an analytic discourse [17]. Thereby, Visual Analytics helps the user gaining insight in the used algorithms and methods. In detail, the collaboration between the human and the computer is most important in our application in the analysis step, where the human's abilities to interpret and assess the results are in demand. Based on that, several work has been done in recent years. Combined with visualizations Oelke and Keim [18] showed in 2007 a new method for Visual Literary Analysis, which is called *Literature Fingerprinting*. The fingerprints are pixel-based visualizations, encoded with colour to show the text features. Tag clouds or word clouds have become more and more in use through the development and applications on the internet. These frugal text visualizations map the word frequency on font size [19]. The success of tag clouds in recent years is due to the fact, that users were allowed to create word clouds with their own content. One of the most famous single-purpose tool for example is wordle [20]. A more general visualization sharing site for example is Many Eyes [21]. It was generally created for explorative data analysis. Wordle is also able to support non-experts to visualize and arrange personally meaningful information [22]. A possibility to enrich word clouds with more information showed Wanner et al. [23]. POSvis [24] is an example for Literature Analysis using a tag cloud amongst others. The authors tried to analyze the book *The Making of Americans*. According to a specialist, the postmodern writing is very hard to read. The various visualizations (bar chart, text snippets) are arranged around a part-of-speech word cloud on a dashboard. Additionally, the software allows the user to explore and analyze the document. We are also use such visualization techniques and give the user the possibility to explore and detect interesting parts of the book.

III. Features To Measure Age Suitability

As mentioned in the previous section, we could identify five different aspects of age suitability in our interviews with booksellers. For each of these properties we separately define a measure to approximate them computationally.

A. Linguistic Complexity Feature

Linguistic complexity can be measured in terms of the vocabulary that is used or with respect to the ease of reading. Measures of vocabulary richness are mainly based on the evaluation of the number of different types (unique vocabulary items) and the overall number of tokens (any occurrence of a word type, i.e. the text length). In this work, we make use of the *Simpson's Index (D)* [14] that calculates the probability that two arbitrarily chosen words belong to the same type.

$$D = \frac{\sum_{r=1}^{\infty} r(r-1)V_r}{N(N-1)}$$

In the formula, N denotes the number of tokens (i.e. the text length) and V_r the number of vocabulary items that occur exactly r times.

To assess the readability of the text, the *Automatic Readability Index*[25], a popular readability measure, is used. It consists of two parts: (a) an estimation of the difficulty of the words that are used (assuming that longer words are more difficult to use) and (b) the average sentence length as an indicator for the difficulty to process the sentence.

$$ARI = 4.71 \cdot (\frac{\#characters}{\#words}) + 0.5 \cdot (\frac{\#words}{\#sentences}) - 21.43^{1}$$

The measure is normalized in a way that the resulting values range between 1 and 12, reflecting the US grade level that is needed to understand the text.

B. Story Complexity Feature

Measuring the complexity of a text on a statistical and syntactic level is reasonable and important, however, there are more factors that contribute to complexity. Next, we are going to look at the discourse level of the text by assessing the complexity of the story line. Measuring text properties on a higher linguistic level than the statistical level is challenging. Usually, there is no way to measure these aspects directly.

[1] # denotes "number of"

provide valuable information about the age group that a book was designed for. The necessary data can be retrieved from online databases.

IV. Evaluation

The evaluation of the different features is done separately. The following two sub-sections handle the Story Complexity and the Topic Detection. The Readability will not be evaluated because no new algorithm was implemented.

The fact that our data consist of whole books make it impossible to get objective ground-truth data. Publisher suggest a minimum age for every book but are perhaps influenced by economic reasons. That is why it was necessary to generate our own ground-truth data. Therefore we used a so called Human Intelligence Task (HIT) with the Amazon Mechanical Turk Service. This service provides a crowd-sourcing marketplace to execute different types of tasks by ordinary people. A single HIT is an online job which can be executed by every Amazon Mechanical Turk member fulfilling the requirements. Our HIT consists of a questionnaire with 14 questions about 15 different books. At least the questions to one book must be answered to receive a small award. Every answer was checked for trustworthiness examining an implemented time stamp and the correlation between two test questions. About 300 questionnaires were answered trustworthy and provide our ground-truth data. Only six of the 15 books were answered often enough to be analysed to guarantee the methodological correctness.

A. Evaluation of Story Complexity

For the evaluation we took the book *Harry Potter and the Philosopher's Stone* with a total of 179 Characters. Following you can see our results:

TABLE I
Results of evaluation

	Relevant	Non Relevant
Retrieved	69	29
Not Retrieved	47	34
Total	116	63

The precision of the algorithm is 0.704 and the recall 0.595. When we took a look in our results we recognized, that the NER process is not consistent over the book. So "Hagrid", a character of the Harry Potter series, is tagged as *person* and elsewhere in the book as an *organization*. If the right tagged noun is never at least once followed or preceded by a communication verb, but the wrong one does so our result gets worse. A solution could be implementing a threshold, e.g. as a hypothesis "Hagrid" is detected 75 percent as a *person* and 25 percent as *organization* then we could assume that "Hagrid" is a person. Although, that could lead to problems (e.g. "Washington") an improvement could be achieved. We would like to try that in the future.

B. Evaluation of Topic Detection

Our implemented algorithm to compute the possibility that a certain book belongs to a specific topic will be evaluated using the answers of the online questionnaire as our ground-truth data. The participants had to choose whether the book is about one or more of the six predefined topics or not. To compare our algorithm with the user opinion the results were normalized between 0 and 1. Additionally the significance value used in our algorithm is examined. Each book is therefore analyzed twice once with the significance value and another time without. The following figure illustrates the evaluation with four different books (Fig. 1).

The bar charts illustrate that the user tendency is much more similar to the algorithm with the significance value than without. However there are exceptions like the book *1984* (bottom left) where both results are misleading. The main part to improve the algorithm are the predefined hardcoded lists of representing words for each topic. With the lists being more complete and correct the whole algorithm performs better.

V. Visual Book Analysis

With the measures that were defined in section III we are able to approximate the different aspects of age suitability computationally. However, it is unclear how much each feature contributes to the overall rating. Furthermore, for some features we do not have a single score but a whole bunch of information that requires interpretation. We therefore decided to make use of visual analysis techniques in the next step of the analysis process. This comes with the following advantages:

- The human visual system is very powerful allowing the user to grasp a large amount of data at an instance as long as it is meaningfully displayed. [32] Visualization therefore is an ideal means of integrating the user into the process.
- Thus, using visualization allows us to provide the detailed information of our measures to the user without causing too much cognitive load.
- It is known that humans are very proficient in detecting visual patterns, a capability that is highly desirable in this case because of the complex measures that are used. With this, the interpretation of the data that is needed to overcome the semantic gap can be left to the human analyst.
- At the same time this comes with the advantage that the human analyst does not need to trust a "black box" but is able to comprehend the decision of the algorithm. This is especially important for features that may be weighted differently depending on the personality of the reader.

In the following, we are going to introduce our visual analysis tool. As the emotion detection and the analysis of the story complexity are the two features that profit most from the visual analysis, their visualizations are presented in detail in sections V-A and V-B. This is followed by a presentation of the full application.

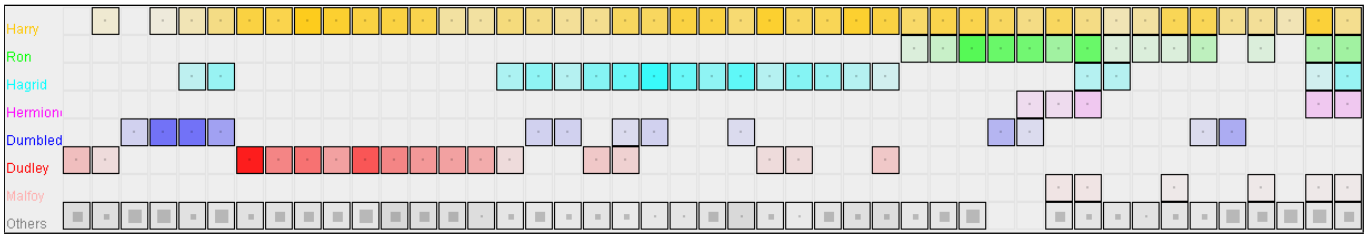

Fig. 2. Story Complexity Visualization of the book *Harry Potter and the Philosopher's Stone*

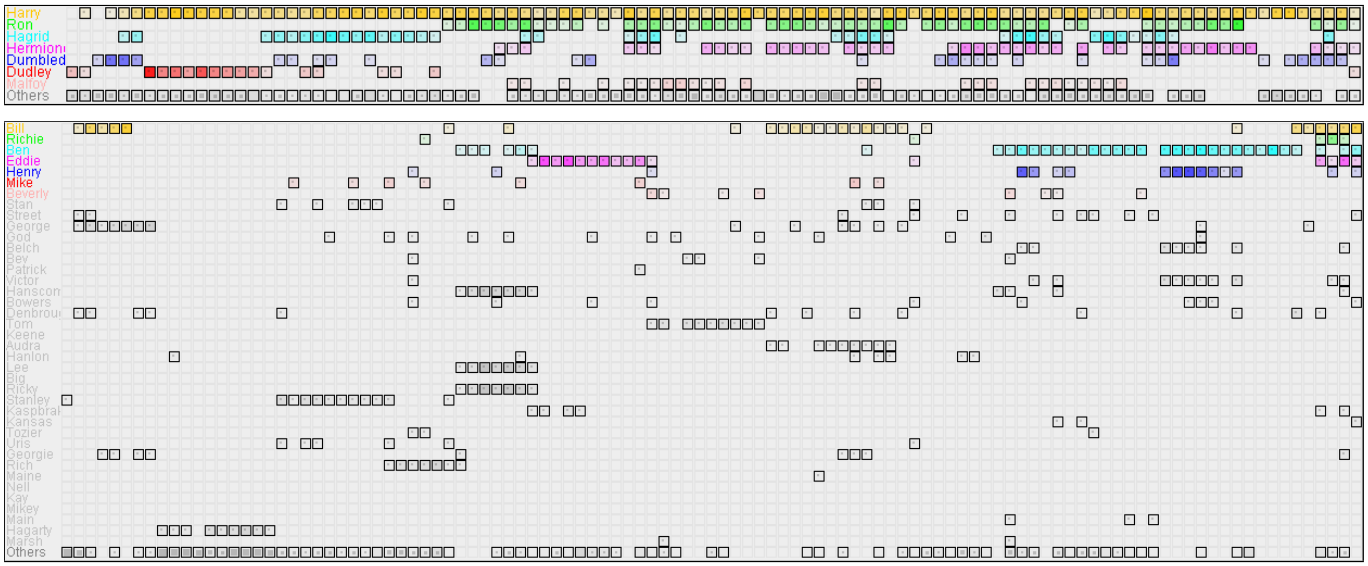

Fig. 3. Comparison of the Story Complexity Visualization of the books *Harry Potter and the Philosopher's Stone* (top) and *It* (bottom).

B. Visualization of Emotions

The four different emotions happiness, sadness, anger and anxiety are visualized in a bar chart diagram. The height of each bar represents the number of detected emotion words for the specific category.

Especially with the emotion feature we are facing the challenge that we need to overcome a gap between what we measure and what we would like to approximate on a semantic level. Remember that we are interested in the *aroused* emotions but can only work with a measure that is based on word associations that are related to emotional states. Thus, an inspection and interpretation of the result by a human expert is critical. We therefore do not only visualize the overall emotion scores, but again calculate separate values for each text unit as for the story complexity. This also gives us the chance to analyze the development of the emotions across the text.

Fig. 4 shows the course of the emotion feature for the book *A Long Way Down*. This detailed view reveals much information about the story. While happiness is the most dominant emotion in most of the book, there is a passage in the middle in which it almost completely disappears. Furthermore, there are several text units in which sadness and happiness (red and yellow bars) occur with a similar strength suggesting that this might be an emotionally demanding part of the book in which the two contrasting emotions are close together. However, at the end of the story the happiness value is clearly dominating which hints at a happy end. Emotion words related to anger are nearly not present at all whereas anxiety is present at a certain level almost all over the book. To investigate a single bar chart in detail, it is possible to display a word cloud of the underlying emotion words (see figure 4).

C. Visual Agesuitability Tool

The final Visual Agesuitability Tool combines the visual representations of the five features in one multi-view dashboard display (see figure 5).

In the upper left corner, a summary of the detected emotions is presented in a bar chart diagram. Users can interactively drill-down to the detailed representation that is presented in section V-B. Similarly, the character panel at the bottom shows an overview representation of the active characters which can be zoomed in to get the in-depth information that is provided by the summary report visualizations that are depicted in figures 2 and 3. Numeric information such as the readability scores, the vocabulary richness, the number of pages, or the number of words per page are shown in the upper middle of the panel. Additionally, color is used to visually encode the numbers and support the user in assessing how these values

REFERENCES

[1] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[2] J. Kadhim and V. Crittenden, "Amazon Mechanical Turk," retrieved from Citeseer.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, pp. 1–135, January 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1454711.1454712

[4] J. Zhang, Y. Kawai, T. Kumamoto, and K. Tanaka, "A novel visualization method for distinction of web news sentiment," in *Web Information Systems Engineering - WISE 2009*, ser. Lecture Notes in Computer Science, G. Vossen, D. Long, and J. Yu, Eds. Springer Berlin / Heidelberg, 2009, vol. 5802, pp. 181–194.

[5] M. L. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner, "User-directed sentiment analysis: visualizing the affective content of documents," in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, ser. SST '06, 2006, pp. 23–30.

[6] R. Nallapati, "Semantic language models for topic detection and tracking," in *NAACLstudent '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 1–6.

[7] S. Green, "Building hypertext links in newspaper articles using semantic similarity," in *Third Workshop on Applications of Natural Language to Information Systems (NLDB'97)*, 1997, pp. 178–190.

[8] S. Scott and S. Matwin, "Text classification using WordNet hypernyms," in *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, 1998, pp. 38–44.

[9] A. Hotho, S. Staab, and G. Stumme, "Wordnet improves text document clustering," in *Proc. of the SIGIR 2003 Semantic Web Workshop*. Citeseer, 2003, pp. 541–544.

[10] J. Allan, Ed., *Topic detection and tracking: event-based information organization*. Norwell, MA, USA: Kluwer Academic Publishers, 2002.

[11] D. Oelke, D. Spretke, A. Stoffel, and D. A. Keim, "Visual readability analysis: How to make your writings easier to read," in *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST '10)*, 2010.

[12] R. Gunning, *The technique of clear writing*. McGraw-Hill, 1952.

[13] J. P. Kincaid, R. P. Fishburn, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas for Navy Enlisted Personnel," Naval Air Station Memphis, Research Branch Report 8-75, 1975.

[14] D. I. Holmes, "Authorship Attribution," *Computers and the Humanities*, vol. 28, pp. 87–106, 1994.

[15] D. Hoover, "Another perspective on vocabulary richness," *Computers and the Humanities*, vol. 37, pp. 151–178, 2003.

[16] D. Oelke, "Visual document analysis: Towards a semantic analysis of large document collections," Ph.D. dissertation, University of Konstanz, 2010.

[17] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, *Mastering the information age—solving problems with visual analytics*. Eurographics Association, 2010.

[18] D. A. Keim and D. Oelke, "Literature fingerprinting: A new method for visual literary analysis," in *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology (VAST '07)*. IEEE Computer Society, 2007, pp. 115–122.

[19] F. B. Viégas and M. Wattenberg, "Timelines tag clouds and the case for vernacular visualization," *interactions*, vol. 15, no. 4, pp. 49–52, 2008.

[20] "wordle, http://www.wordle.net/, october 31st, 2010."

[21] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon, "Manyeyes: a site for visualization at internet scale," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, pp. 1121–1128, 2007.

[22] F. B. Viégas, M. Wattenberg, and J. Feinberg, "Participatory visualization with wordle." *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1137–1144, 2009.

[23] F. Wanner, M. Schaefer, F. Leitner-Fischer, F. Zintgraf, M. Atkinson, and D. A. Keim, "Dynevi - dynamic news entity visualization," in *Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010)*, J. Kohlhammer and D. A. Keim, Eds., Jun. 2010, pp. 69–74.

[24] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar, "What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections," in *IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST)*, Oct. 2009, pp. 107–114. [Online]. Available: http://liris.cnrs.fr/publis/?id=4360

[25] R. Senter and E. Smith, "Automated Readability Index," 1997, technical Report.

[26] D. Klein, J. Smarr, H. Nguyen, and C. Manning, "Named entity recognition with character-level models," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, 2003, pp. 180–183.

[27] Ubiquitous Knowledge Processing (UKP) Lab, TU Darmstadt, English communication verbs, www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/data/english_communication_verbs.txt.

[28] H. L. Chieu and H. T. Ng, "Named entity recognition: a maximum entropy approach using global information," in *Proceedings of the 19th international conference on Computational linguistics*, 2002, pp. 1–7.

[29] C. Fellbaum, *WordNet: An electronic lexical database*. MIT Press, 1998.

[30] "Frequency list from the brown corpus, www.edict.com.hk/lexiconindex/frequencylists/words2000.htm."

[31] C. John, "Emotionality ratings and free-association norms of 240 emotional and non-emotional words," *Cognition & Emotion*, vol. 2, no. 1, pp. 49–70, 1988.

[32] C. Ware, *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, 2004.

[33] D. Oelke, M. Hao, C. Rohrdantz, D. Keim, U. Dayal, L. Haug, and H. Janetzko, "Visual opinion analysis of customer feedback data," in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE, 2009, pp. 187–194.