Enriching Information Retrieval with Reading Level Prediction

Kevyn Collins-Thompson
Microsoft Research
1 Microsoft Way
Redmond, WA USA 98052
kevynct@microsoft.com

ABSTRACT

The ability of a user to understand a document would seem to be an critical aspect of that document's relevance, and yet a document's reading difficulty is a factor that has typically been ignored in information retrieval systems. In this position paper we advocate for incorporating estimates of reading proficiency of users, and reading difficulty of documents, into retrieval models, representations for learning algorithms, and large-scale analyses of information retrieval systems and users, particularly for Web search. We describe key research problems such as estimating user proficiency, estimating document difficulty, and re-ranking, and summarize some potential future extensions that could exploit this new type of meta-data.

1. INTRODUCTION

Traditionally, search engines have ignored the reading difficulty of documents and reading proficiency of users in modeling relevance, ranking documents, and many other aspects of retrieval. This is especially evident with the increased interest in more effective search systems for children and students [11]. While addressing children's search needs requires solving many important problems in interface design, content filtering, and results presentation, one fundamental problem is simply that of providing relevant results at the right level of reading difficulty. Similarly, experts may not want tutorials and introductory texts and instead prefer material that is actually highly technical. Non-native language speakers also form a significant population of users who could benefit from improvements in information retrieval that account for reading level.

We propose using estimates of reading proficiency for users, and reading difficulty for documents, to improve system usability and relevance for broad classes of users and tasks: from user profiles, models of session and query intent, ranking algorithms, document classifiers, and summarization or result presentation algorithms. To date, however, there has been little, if any, published work on user modelling and re-ranking algorithms based on reading level and their deployment and evaluation.

A simple example in Table 1 shows why reading level can be important to model for ranking search results. Three out of the top four results for the query `[insect diet]` are at a 'high' level of difficulty according to a number of readability measures[1]. These results may be appropriate for a high school student or academic, but would be less appropriate for, say, a third-grade elementary school student. This problem is representative of the large potential mismatches in user vs. document level we have observed across many other Web queries.

2. KEY RESEARCH PROBLEMS

There are several core areas of research we are currently exploring that are central to incorporating reading level meta-data into retrieval systems in a useful way.

2.1 Understanding the reading level properties of the Web and its users

Currently little is known about basic reading level-related properties of the Web, or the nature of user interactions and queries with respect to reading difficulty. Thus, there is a need for large-scale reading-level analysis of the Web that examines properties like the relationship of reading level meta-data to other meta-data for the same pages, such as ODP category distributions [3]; analysis of differences in reading level distributions across different domains and types of pages, such as high- versus low-traffic pages; and interesting hyperlink-based clusters with low and high inter-page differences in level. Some recent work has begun to study user interactions via query logs: Torres et al. [12] performed an analysis of the AOL query log to characterize so-called 'Kids' queries. A query was labeled as a Kids query if and only if it had a corresponding clicked document whose domain was listed as an ODP entry in the 'Kids&Teens' ODP top-level category. Other work has explored query expansion methods for queries formulated by children [13]. More analysis is needed to obtain a better understanding of where and how reading-level meta-data is likely to be most effective for specific search tasks or groups of users.

2.2 Personalizing search by reading difficulty

There are three key problems to solve in order to incorporate reading level as a relevance signal for personalized Web search: estimating reading level of documents; estimating

SIGIR 2011 Workshop on Enriching Information Retrieval (ENIR 2011)
July 28, 2011, Beijing, China.
.

[1] For this example we used a variant of the method published in [6], but two other standard measures applied to the full page text were consistent with that prediction.

reading proficiency of users; and ranking documents based on reading level of users and documents.

2.2.1 Estimating reading difficulty of documents

Estimating reading difficulty has been studied for decades, but traditional formulae such as Flesch-Kincaid provide only a crude combination of vocabulary and syntactic difficulty estimates. Recent progress has been made in applying statistical modeling and machine learning to improve general-purpose reading difficulty estimation for non-traditional documents [6][8][10] such as Web pages or short snippets. Current reading level prediction algorithms are based on supervised machine learning, using vocabulary and syntax features extracted from labelled data, where the labels typically correspond to school grade levels such as American grade levels 1 through 12. Because of the many factors that can influence comprehension, reading level prediction is an imprecise task, and current state-of-the-art prediction accuracy is approaching that of human judges. Statistical methods produce a posterior distribution over grade levels to capture uncertainty in the prediction, which is useful in improving the reliability of using this meta-data in ranking and retrieval.

Using the Web hypergraph is a promising enhancement to prediction. In a preliminary study, Gyllstrom and Moens [7] proposed a binary labeling of Web documents into material for children vs. adults, where the label is inferred using a PageRank-inspired graph walk algorithm called AgeRank. They evaluated this method on a small subset of Web pages. The key advantage is the use of hyperlinks to propagate labels through the Web graph. This approach also included non-vocabulary features such as page color, font size, etc. to help determine the page label. The combination of Web graph, vocabulary, and non-vocabulary features with existing machine learning methods is likely to provide a good basis for reading level meta-data of documents.

2.2.2 Estimating reading proficiency of users

One approach is to have users self-identify their level of proficiency. This is the approach Google has used in their recent deployment of an Advanced Search feature to filter results by Low, Medium, and High levels of difficulty. However, self-identified user information may not always be available or reliable, in which case we need ways to construct a reading proficiency profile automatically. To our knowledge there has been little work on automatically estimating a reading proficiency profile for a specific user. We expect that existing learning algorithms could be applied based on such observations as the reading level of past (satisfied) clicked documents; semantic or syntactic features of current and past queries; or previously visited pages or domains from a known list of expert or kids'-related sites, and other features of the user's history or behavior. More generally, we also forsee the need for models that capture *expertise* on specific topics, in addition to general reading proficiency.

2.2.3 Re-ranking based on reading difficulty

Re-ranking using reading level aims at reducing the 'gap' between the user's reading proficiency distribution and a document's reading level distribution. Interestingly, with preliminary prototypes we have found that the ability to re-rank the results of a high-difficulty technical query using a low-difficulty user model can be very useful in finding tutorial or introductory material. This re-ranking tended to demote Wikipedia articles that matched the query well but had more dense technical vocabulary, while promoting blog entries where the authors were explaining the same technical concepts to their readers using more colloquial language. As with other types of personalization there is a risk-reward tradeoff: we want to promote documents closer to the user's reading proficiency level, while not straying too far from the default ranking, which is typically a highly-tuned relevance signal optimized for the 'average' user. Exploring reliable methods for modifying existing rankings based on a user reading proficiency profile is an area of current research. Another important open problem is how to combine reading level personalization with other types of personalization based on location, topic or other meta-data.

3. FUTURE EXTENSIONS

Beyond the core problems presented above, reading level prediction and user proficiency profiles may be applied in a variety of other ways in search systems. We briefly describe two directions here.

3.1 Influencing retrieval presentation

Beyond ranking, there are other aspects of retrieval that could benefit from estimates of reading level. The captions or snippets of search results could be tailored to focus on vocabulary more familiar to the user. There is some existing body of research on automatic text simplification [9][4] that could be augmented to produce summarizations with personalized knowledge of which words a user knows or doesn't know based on their reading proficiency profile. Query suggestions could include custom expansion terms or reformulations that reflect special learner-centric or expert-centric intent. The layout simplicity, color, font size, and other aspects of display could be likewise be adjusted.

3.2 Adapting documents and users to each other

Instead of assuming the reading level of users and documents is something to be passively observed, we can propose a new class of algorithms that actively *adapt* document or user knowledge in order to reduce the 'knowledge gap' between them when a mismatch occurs. For example, when returning a search result whose difficulty is higher than the user's current proficiency, the system could identify important *words to learn* in the results that the user is not likely to know – e.g. a search on articles about *stomach aches* might return pages that also use the technical term *gastritis*. In such cases, the system could provide links to supporting definitions or background material[2]. Such algorithms would need to be able to identify key vocabulary in a document; compare against a user's reading proficiency model; and compute the best small subset of critical 'stretch' vocabulary required to understand most of a document. Other relevant scenarios include intelligent tutoring applications that help stretch the student's vocabulary by retrieving content that is slightly above their current reading level, along with satisfying other linguistic properties that align with curriculum goals [5]. In a related direction, Agrawal et al. use estimates of syntactic complexity and key concepts to identify difficult sections of textbooks that could benefit from better exposition [1] and to find links to authoritative content [2]. The educational potential for such augmentations, especially those based on individual user models, seems very promising.

[2] Thanks to Jaime Teevan for this suggestion.

Rank	URL Domain	Title	Category	Reading Level (Grade level)
1	insectdiets.com	Insect Diet & Rearing Research	Technical/Research	High (10.0)
2	imfc.cfl.rncan.gc.ca	Insect Diet	Technical/Research	High (10.0)
3	www.sugar-glider-store.com	Insect-Eater Diet	Commercial	Medium (7.0)
4	insectdiets.com	Insect Rearing Research	Technical/Research	High (10.0)
5	insectrearing.com	Bio-Serv Entomology Division	Commercial, Technical	Medium (8.0)
6	www.ehow.com	Aquatic Insects & Diet	Educational	Medium (7.0)
7	www.exoticnutrition.com	Insect-eater Diet...	Commercial	Medium (6.0)
8	www.tutorvista.com	Insect diet: Questions & Answers	Educational	Low (5.0)
9	www.encarta.msn.com	Dictionary	Not relevant (empty)	N/A
10	deltafarmpress.com	Producers may put fish on insect diet	Technical/News	High (10.0)

Table 1: Top ten results, in rank order, for the query [*insect diet*] from a commercial search engine, showing the wide variation in reading level that can occur for material retrieved on the same query. Reading level here is estimated using a statistical model [6] and shown in brackets. (Query issued on January 20, 2011.)

4. CONCLUSION

Research in applying meta-data derived from reading level prediction to the Web and other information retrieval domains is only just beginning, and we believe it has the potential to improve the performance of a wide range of retrieval tasks for individual users: from personalized Web search to educational applications.

5. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Identifying enrichment candidates in textbooks. In *Proceedings of WWW 2011*, pages 483–492.

[2] R. Agrawal, S. Gollapudi, K. Kenthapadi, N. Srivastava, and R. Velu. Enriching textbooks through data mining. In *ACM Symposium on Computing for Development (ACM DEV).*

[3] P. Bennett, K. Svore, and S. Dumais. Classification-enhanced ranking. In *Proceedings of the 19th Annual International World Wide Web Conference (WWW '10)*, pages 111–120, 2010.

[4] J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. Simplifying text for language-impaired readers. In *Proceedings of EACL '99*, pages 269–270.

[5] K. Collins-Thompson and J. Callan. Information retrieval for language tutoring: an overview of the REAP project. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–545, 2003.

[6] K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT-NAACL 2004*, pages 193–200.

[7] K. Gyllstrom and M.-F. Moens. Wisdom of the ages: Toward delivering the children's web with the link-based Agerank algorithm. In *Proceedings of CIKM 2008*, pages 159–168.

[8] P. Kidwell, G. Lebanon, and K. Collins-Thompson. Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP) 2009*, pages 900–909.

[9] C. Napoles and M. Dredze. Learning Simple Wikipedia: A cogitation in ascertaining abecedarian language. In *Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids at NAACL-HLT 2010.*

[10] S. E. Petersen and M. Ostendorf. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(1):89–106.

[11] PuppyIR. PuppyIR: An open source environment to construct information services for children. 2011. http://www.puppyir.eu/.

[12] S. D. Torres, D. Hiemstra, and P. Serdyukov. An analysis of queries intended to search information for children. In *IIiX 2010*, pages 235–244.

[13] M. van Kalsbeek, J. de Wit, D. Trieschnigg, P. van der Vet, T. Huibers, and D. Hiemstra. Automatic reformulation of children's search queries. Technical Report TR-CTIT-10-23, June 2010.