Automatic Recognition of Reading Levels from User Queries

Xiaoyong Liu, W. Bruce Croft, Paul Oh, and David Hart
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
xliu@cs.umass.edu

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query formulation*

General Terms: Experimentation

Keywords: Query Inference, Reading Level, Readability, Query Classification, Context, Answer Level, Personalization

1. INTRODUCTION

In general, interactions with Web search engines could be characterized as "one size fits all". There is no representation of user preferences, search context, or the task context. Because of this obvious limitation of current search technology, personalization and context have been identified as major research challenges in a number of workshops. This work describes an ongoing effort toward automatically inferring knowledge about users and context from user queries. We report experiments that focus on the recognition of one context feature, namely reading level, and show that queries from users of different level groups can be effectively separated.

Scientific readablility [4] has been studied for decades. Many readability indices have been developed to determine the reading levels of texts. Popular ones include Flesch-Kincaid [3], SMOG [5], and FOG [2] tests. A statistical approach has been proposed recently in [6]. A common characteristic among all these methods is that they have been developed for documents or passages that have at least 100 words or 10 sentences. They become unreliable when the size of the text drops below the requirement, which is typical for user queries. The originality of this work lies in the development of a method for identifying reading levels based solely on the very short query texts. We treat readability estimation as a text categorization problem and employ Support Vector Machine learning algorithms based on both syntactic and semantic features. The proposed method can be used directly in an information retrieval or question-answering system to determine the level of answers appropriate for the users.

2. DATA

The data for this study come from 3 different sources. The queries for the category K-6 are collected from a local elementary school during the time period of June 16-30, 2003. They are questions raised by students from grade K-6 about various topics discussed in science classes or related assignments. Queries for the category Excite are taken randomly from the log of queries submitted to the Excite search engine on Dec. 20, 1999. Queries for the remaining categories are archived questions from 1996 to 2002 of the Mad Scientist question-answering service [7]. We cleaned the data by removing queries that do not contain any alphanumeric letters or those that only contain phrases like "thank you", "see below", etc. The statistics of the resulting data set are summarized in table 1.

Table 1. Statistics of grade level data.

Grade level	# of queries	Avg. # of words per query	Avg. # of characters per word	Avg. # of syllables per word
K-6	407	6.93	4.22	1.36
7-9	2508	9.13	4.64	1.51
10-12	3374	9.20	4.78	1.57
Undergrad	2414	9.23	4.87	1.60
Grad	1669	9.24	4.85	1.59
Excite	1999	3.35	5.86	1.83

3. PILOT STUDY

In order to establish the merit of automatically recognizing reading levels from user queries in helping retrieval, we carry out a pilot study that examines whether grade relevant materials can be retrieved if we have knowledge about users' reading levels. A set of 40 grade K-6 queries and 60 grade 7-9 queries are randomly selected from the data reported in table 1. We submitted these queries to the Google search engine on Feb. 9, 2004, and saved the top 20 retrieved Web pages by the engine. Only the entry pages were saved. The pages that were linked to the entry pages were not used. A local school teacher who is familiar with both elementary and middle school science classes is presented with the queries and retrieved Web pages and asked to judge whether each of the Web pages is topically relevant and/or grade relevant. A Web page is considered topically relevant if it provides an answer and in a similar context to that of the query. For example, for the query "What ocean is deepest?", if a Web page describes Pacific Ocean but does not mention that it is the deepest ocean, we do not consider the Web page to be topically relevant even though "Pacific Ocean" is the correct answer. If a topically relevant page is written in a way that can be comprehended by an average user from the grade category under consideration, it is labeled grade relevant. Out of the 40 K-6 queries, it was found that 11 of them had no relevant pages within top 20 retrieved, 9 queries had topically relevant pages but no grade relevant ones, and 20 queries had both topically and grade relevant pages. For the 60 queries from grade 7-9, 28 of them had no relevant pages within top 20 retrieved, 8 queries had topically relevant pages but no grade relevant ones, and 24 queries had both topically and grade relevant pages. We report in table 2 the average number of topically and grade relevant pages retrieved for those queries that had both and the average rank at which the first topically and grade relevant pages are retrieved respectively.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

Table 2. Statistics of topically and grade relevant Web pages within top 20 retrieved per query.

Cat.	Avg. # of topically rel. pages	Avg. rank of 1st topically rel. page retr.	Avg. # of grade rel. pages	Avg. rank of 1st grade rel. page retr.
K-6	8.3	1.8	2.9	6.4
7-9	6.9	2.4	3.4	4.4

A couple of observations can be made. The first and obvious is that current search technology to handle natural language queries is still very limited. The second and more important one is that, when relevance has to depend on both content and level which is usually the case for a Web search, for about 50% of the queries for each grade category retrieval performance can be improved by displaying grade appropriate pages before other pages.

4. EXPERIMENTS AND RESULTS

We use Support Vector Machines (SVMs) to learn the differences between and classify queries from different grade categories. The experiments are done using the LIBSVM[1] software package. A number of syntactic and semantic features are derived from queries. Examples of syntactic features include sentence length, average number of characters per word, average number of syllables per word, percentage of various part-of-speech tags, and various readability indices such as Flesch-Kincaid, SMOG, and FOG. Semantic features include frequency of numerous 1-, 2-, and 3-word sequences. We make a random split of the data for each category into 90% training and 10% test instances. For experiments involving two or three categories, we combine the training instances from those categories to form the training set and similarly for the test set. Model selection is done by 5-fold cross validation with exhaustive parameter search on the training data. The best parameter combination is then applied on the test data. In our experiments, radial kernel gives the best performance.

We first compare the recognition accuracy of reading levels from queries involving two categories by our SVM-based approach with that by standard readability indices. The reading level of the queries from the Excite category cannot be clearly defined by readability indices and hence are excluded from this experiment. Results are shown in table 3. We observe that the readability indices perform poorly for all 2-category combinations whereas the SVM-based approach perform significantly better and can effectively distinguish queries from different categories.

Table 3. Comparison of recognition accuracy(%) between readability indices and our SVM-based approach on test data.

Categories	Flesch-Kincaid	SMOG	FOG	SVM-based
K-6+7-9	40.0000	13.7931	14.1379	93.4483
K-6+10-12	24.9337	10.6101	10.6101	95.7560
K-6+undergrad	19.5730	14.2349	14.2349	86.1210
K-6+grad	20.8739	19.4175	19.4175	92.7184
10-12+grad	11.5308	0	0	66.6004

In the second experiment, we examine whether queries about science (e.g. from K-6, 7-9, 10-12, and Grad) can be distinguished from general adult-level search engine queries (e.g. Excite). Results are reported in table 4. We observe that a good separation (over 83% accuracy) can be obtained by the SVM-based approach.

The next experiment combines queries from 3 different categories and further evaluates the performance of our approach. Results are also given in table 4. Again, the SVM-based approach is found to perform reasonably well for different 3-category combinations. A recognition accuracy close to or above 80% can be achieved.

Table 4. Reading level recognition accuracy (%) for 2-category and 3-category cases using our SVM-based approach.

Type	Categories	Accuracy on training data	Accuracy on test data
2-cat	K-6+7-9	91.2762	93.4483
	K-6+10-12	93.9483	95.7560
	K-6+undergrad	85.5512	86.1210
	K-6+grad	91.8717	92.7184
	10-12+grad	66.9163	66.6004
	K-6+excite	96.3544	97.0711
	7-9+excite	90.9315	89.5323
	10-12+excite	91.3376	92.1642
	Grad+excite	84.0145	83.5616
3-cat	K-6+7-9+excite	86.0565	86.2986
	K-6+10-12+excite	88.2782	89.5833
	K-6+undergrad+excite	80.0000	78.5417
	K-6+grad+excite	81.9753	81.7284

5. CONCLUSIONS AND FUTURE WORK

We proposed and evaluated an SVM-based approach to automatically recognizing reading levels from user queries. Early results show that the proposed method performs significantly better than standard readability indices and it can achieve a recognition accuracy close to or well above 80% for both 2-category and 3-category cases. A natural extension of this work is to incorporate it into retrieval. We have shown through the pilot study that potential improvement in retrieval performance (especially in a Web search) can be achieved by matching the query and documents not only by content but also by level. We have taken the first step by inferring levels from queries. The next steps would be applying a method similar to what we proposed here to determine the reading levels of documents, and incorporate a representation of context information (reading levels in this case) into the retrieval model.

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant number DUE-0226144.

7. REFERENCES

[1] Chang. C. & Lin, C. (2001). LIBSVM: a library for support vector machines. Software is available for download at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[2] Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill.

[3] Kincaid, J.P., Fishburn, R.P., Jr. Rogers, R.L., and Chissom, B.S. (1975). Derivation of new readability formulas for Navy enlisted personnel. *Research Branch Report 8-75*, Naval Air Station Memphis, Millington, Tennessee, 40 pages.

[4] Klare, G. R. (1963). *The Measurement of Readability*. Iowa State University Press.

[5] McLaughlin, H. (1969). SMOG grading - a new readability formula, *Journal of Reading*, 22, 639-646.

[6] Si, L. & Callan, J. (2001). A statistical model for scientific readability. In *CIKM'01 Proceedings*.

[7] http://www.madsci.org/