An Unsupervised Technical Difficulty Ranking Model Based on Conceptual Terrain in the Latent Space[*]

Shoaib Jameel Wai Lam Xiaojun Qian
Department of Systems Engineering and
Engineering Management
The Chinese University of Hong Kong
{msjameel,wlam,xjqian}@se.cuhk.edu.hk

Ching-man Au Yeung
ASTRI
3/F, Bio-Informatics Centre, Science Park
Hong Kong
albertauyeung@astri.org

ABSTRACT

Search results of the existing general-purpose search engines usually do not satisfy domain-specific information retrieval tasks as there is a mis-match between the technical expertise of a user and the results returned by the search engine. In this paper, we investigate the problem of ranking domain-specific documents based on the technical difficulty. We propose an unsupervised conceptual terrain model using Latent Semantic Indexing (LSI) for re-ranking search results obtained from a similarity based search system. We connect the sequences of terms under the latent space by the semantic distance between the terms and compute the traversal cost for a document indicating the technical difficulty. Our experiments on a domain-specific corpus demonstrate the efficacy of our method.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering, Retrieval Models, Search Process, Selection Process; H.3.7 [**Digital libraries**]

General Terms

Algorithms, Experimentation

Keywords

LSI, Terrain, Conceptual Difficulty, Ranking

1. INTRODUCTION

When it comes to searching for domain-specific content on the web, people with diverse background query search engines to locate the text which is both relevant and can fit the level of understanding. General web search engines are too broad in terms of topical coverage. Simply using the link structure such as PageRank [4] or simple cosine similarity [7] based measure to retrieve and rank documents will not achieve the goal of matching conceptual difficulty. Technical expertise varies from one person to another and it is difficult to cater for such needs without building a user model [8] for every user and reflect results accordingly.

We present an unsupervised method for re-ranking domain-specific documents given a query based on the technical difficulty of documents. We first obtain search results for a query from an IR system. After that the retrieved documents are re-ranked automatically based on the conceptual difficulty using our proposed model.

2. THE CONCEPTUAL TERRAIN MODEL

We use LSI for our task because it exploits the co-occurrence matrix to bring out new structural relationships between the terms and documents in the latent space. Terms which contribute with more synergy in one context will be *close* to that context in the latent space [1]. In the LSI latent space, our model considers *hop* from one term to another in sequence to determine the conceptual cohesion between the terms in the document.

2.1 Inter-term Cohesion

For the smoothness of conceptual comprehensibility, connections between terms with the surrounding parts in the text are necessary. It is because the interpretation of elements in a discourse is dependent on how the reader is able to semantically relate with other elements in the discourse [2, 3]. In a document d, the inter-term cohesion distance between the term t_n and the term t_{n+1} is denoted by $\hat{s}(t_n, t_{n+1})$ which is computed using Euclidean distance formula in the LSI latent space. We normalize the semantic distance between the two consecutive terms t_n and t_{n+1} by computing the distance between t_n and each term τ which follows t_n in the entire corpus.

$$s(t_n, t_{n+1}) = \frac{\hat{s}(t_n, t_{n+1})}{\sum_{\tau} \hat{s}(t_n, \tau)} \qquad (1)$$

2.2 Term Difficulty

The term difficulty score $\hat{F}(t_n, d)$ for every term t_n in the document d is formulated as:

$$\hat{F}(t_n, d) = idf_{t_n} \times \frac{1}{r(t_n, d) + \epsilon} \qquad (2)$$

idf_{t_n} is the Inverse Document Frequency as in [5]. The Euclidean distance between the term and document vectors in the LSI latent space is denoted by $r(t_n, d)$. ϵ is a very small constant added to accommodate the case when $r(t_n, d) = 0$ and in general $\epsilon << \min(r(t_n, d))$. We normalize $\hat{F}(t_n, d)$ such that its global sum is 1. Let D represent the total

[*] The work is supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: CUHK413510). This work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

number of documents in the corpus. We adopt the following normalization formula:

$$F(t_n, d) = \frac{\hat{F}(t_n, d)}{\sum_{i=1}^{D} \hat{F}(t_n, d_i)} \quad (3)$$

2.3 Traversal Cost in the Conceptual Terrain

The conceptual difficulty is directly proportional to the term difficulty and also inter-term cohesion i.e., inter-term semantic distance. It means that the more the term difficulty and the greater the semantic distance, the more will be the technical difficulty of the sequence of terms. Assume a reader begins from the term t_n in the document d with term difficulty denoted by $F(t_n, d)$ and hops to the next term t_{n+1} in sequence whose term difficulty score is denoted by $F(t_{n+1}, d)$ and covers a semantic distance of $s(t_n, t_{n+1})$ in between the two terms in the latent space. We then compute the traversal cost, $C_d(t_n, t_{n+1})$ for each sequential term bigram in document d, which is expressed as:

$$C_d(t_n, t_{n+1}) = \beta\Big[F(t_n, d) + F(t_{n+1}, d)\Big] + (1-\beta)s(t_n, t_{n+1}) \quad (4)$$

where $0 \leq \beta \leq 1$ is a parameter indicating the role that each of the components plays in determining the technical expertise of a document.

Ranking: Our approach focuses on the relative "technical difficulty" of a document when traversing through the text sequentially, where the difficulty of terms are measured by both the scores in the latent space that represent the deviation from common terms and the cost of transitions from term to term. We aggregated the term difficulties and inter-term cohesion scores to come up with the document's conceptual difficulty score normalized by the document length and then re-ranked the results from simple documents to advanced.

3. EXPERIMENTS AND RESULTS

Existing standard IR test collections do not fulfill our purpose of evaluation as we need domain-specific documents and technical difficulty judgment for the documents. We chose Psychology domain to conduct our experiments and crawled about 170,000 documents from Psychology websites. No term stemming was performed and stop words were not removed. We obtained 110 queries (related to Psychology domain) from AOL query logs. We used Zettair[1] to conduct retrieval and obtained a ranked list using the Okapi BM25 [6] ranking function. We then selected the top ten documents retrieved from the ranked list. We refer our model as "Terrain". The number of latent concepts was 200. The parameter β in Equation 4 was set to 0.5 so that the two components viz. the term difficulty and inter-term cohesion could equally contribute in determining the overall document's conceptual difficulty.

To obtain a ground truth of the technical difficulty of the documents for evaluation purpose, human annotators who were undergraduate students having varied background were invited. They had basic knowledge about Psychology. The annotators were fluent in reading English passages. The standard deviations of judgments among the annotators were 1.23. We compared our method with popular ranking methods and also re-ranked documents using readability formulae. We used NDCG as our evaluation metric.

Table 1: NDCG@10 results. "Terrain" is our proposed model.

Terrain	Cosine	BM25	ARI	Coleman	Flesch	Fog	Kincaid	LIX	SMOG
0.552	0.302	0.347	0.390	0.376	0.365	0.413	0.402	0.390	0.369

Table 1 shows that our terrain model has performed better than the traditional readability formulae and popular ranking functions such as Okapi BM25 and cosine. According to a single-tailed student's t-test our results are statistically significant ($p < 0.05$). What makes our model superior when compared with other models is that we are able to effectively capture term difficulties of domain-specific terms based on their contextual information. It means that in one technical discourse, if a term is used as a general term, its difficulty will be low. However the same term in a more technical thematic discourse will have a high technical difficulty score. Our model also captures conceptual leaps during terrain traversal. The leaps are captured locally with respect to the terms in the vicinity rather than in a global perspective as done by many other models. Thus our terrain model effectively captures inter-term cohesion better.

4. CONCLUSIONS

We have presented our terrain based document re-ranking model based on conceptual difficulty which has outperformed comparative methods. Our approach considers two components in determining the document's technical difficulty which are term difficulty and inter-term cohesion. Traditional readability formulae cannot capture domain-specific jargon, for example, *"star"*, *"shock"* and hence perform poorly on such documents.

5. REFERENCES

[1] J. Bellegarda. A multispan language modeling framework for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6(5):456 –467, sep 1998.

[2] P. Freebody and R. C. Anderson. Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly*, 18(3):pp. 277–294, 1983.

[3] W. Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95:163–182, 1988.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *World Wide Web Internet And Web Information Systems*, (1999-66):1–17, 1998.

[5] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

[6] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *TREC*, 1994.

[7] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[8] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proc. of 14th CIKM*, pages 824–831, 2005.

[1] http://www.seg.rmit.edu.au/zettair/index.html