A Study on Combination of Block Importance and Relevance to Estimate Page Relevance

Shen Huang Yong Yu Shengping Li Gui-Rong Xue Lei Zhang

Department of Computer Science and Engineering Shanghai Jiao Tong University
Shanghai, 200030, P.R.China
+86-21-54745879*603

{shuang, yyu}@cs.sjtu.edu.cn {lishengping, grxue, tozhanglei}@sjtu.edu.cn

ABSTRACT

Some work showed that segmenting web pages into "semantic independent" blocks could help to improve the whole page retrieval. One key and unexplored issue is how to combine the block importance and relevance to a given query. In this poster, we first propose an automatic way to measure block importance to improve retrieval. After that, user information need is also concerned to refine block importance for different users.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Relevance feedback*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia

General Terms

Algorithms, Performance.

Keywords

Block importance, block relevance, iterative combination, information need

1. INTRODUCTION

Traditional retrieval models tend to be affected by the miscellaneous content of the web page, i.e. a page may contain multiple blocks with different contents. Such block information within a web page should be fully exploited. Based on previous page segmentation work [1][2][4], some research efforts have shown that segmenting a web page into several relatively independent blocks can facilitate web search, web link analysis and web mining [7].

We claim that when block information is exploited to estimate the relevance of whole page, one key and unexplored issue is how to combine the block importance and relevance to a given query. Block importance we mentioned here is similar to that defined in [7], i.e. whether a block expresses the major idea of a whole page. To our best knowledge, most block-based whole page retrieval focus on using block information to perform query expansion, or using blocks-based ranking score to revise the original document-based ranking score. Such strategies do not consider the importance of block. Song et al. [7] tried to learn the importance of blocks using both spatial and content features about the page presentational layout. The authors did not present any detailed method and experimental result.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

We think block importance should be measured in both editor's view and web user's view. To the former view, we define *topic coherence* in content between block and whole page. Some noise like navigation bar, copyright will be filtered out. However, some page contains several topics and it's hard to tell which one is the major idea. This is often happen on news page, or a page describing several aspects of a general topic. In such scenarios, the same block will have different importance for users with different information need. We try to model the information need of users and use it to estimate the block importance in user's view. So we propose an iterative way to combine block importance and relevance. The basic idea is enlightened by some work focus on multiple features based iterative combination [3][8]. The importance within a page and relevance to a given query can be considered as two features of a block, which mutual reinforcement each other to improve the retrieval performance. We call such approaches iterative combination.

By these motivations, in this poster, we try to find a marriage of block importance and block relevance, which can improve the search for the web page that contains several topics or noisy information. The major contributions are:

- The general idea of an automatic way is proposed to measure the importance of blocks in webpage editor's view.
- The general idea of an iterative combination is proposed to combine block importance and relevance.
- Two cases studies are presented to show the implements of above general ideas in two retrieval models: language modeling and VSM.

2. GENERAL IDEA OF TOPIC COHERENCE AND ITERATIVE COMBINATION

Before we introduce the general idea of iterative combination of block importance and relevance, the automatically-mined block importance is defined using *topic coherent* between block and whole document.

2.1 Automatic Measurement of Block Importance

VIPS [7] propose how to use spatial and content features to measure a block importance. The content features they use focus on the HTML tags, which are used for the presentational layout. Different with this work, we assume that if a block is a major topic within a page, it will be more coherent with the whole page in topic. Without any training, block importance is estimated automatically as follows:

$$Imp(B_i, D) = \frac{TopicCo(B_i, D)}{\sum_{j=1}^{|B|} TopicCo(B_j, D)}$$

where B_i is the *ith* block, $TopicCo(B_i, D)$ is the topic coherence between block B_i and document D, $Imp(B_i, D)$ is the block importance given the document. After block importance measurement, the relevance of a whole document D can be estimated using linear combination in following way:

$$Rel(D,Q) = \sum_{B_i \in D} Rel(B_i, Q) \times Imp(B_i, D) \quad (1)$$

2.2 Iterative Combination

The measurement of block importance aims to filter out noisy information and help the retrieval of documents. We try to model users' information need and further refine the importance using block or document relevance. The iterative combination idea is a kind of pseudo relevance feedback. What we do here is to find some *important terms* using preliminary retrieval result and gradually refine the block importance using query modeling. Then new block importance will affect the further retrieval. Query modeling can be performed in two ways: using relevant blocks or documents.

$$Q' = \sum_t \sum_{B \in R} Rel(B,t) \frac{Rel(B,Q)}{\sum_{B' \in R} Rel(B',Q)} \quad (2)$$

$$Q' = \sum_t \sum_{B \in R} Rel(B,t) \frac{Rel(D,Q)}{\sum_{D' \in R} Rel(D',Q)} \quad (3)$$

Then the topic coherence between block and query model Q' is used to estimate block importance. And the new relevance is estimated by:

$$Rel'(D,Q) = \sum_{B_i \in D} Rel(B_i, Q) \times (\alpha Imp(B_i, D) + (1-\alpha) Imp(B_i, Q')) \quad (5)$$

3. TWO CASE STUDIES

In this section, we show how to implement above idea in Language Model (LM) for retrieval. Language modeling approach [5] treats each document as a language model and the generation of queries as a random process. Then documents are ranked according to the generation probabilities. Similarly, we use the probability of the block occurring in the web page $\hat{p}(B \mid M_D)$, to measure topic coherence between a block and a whole page. The assumption behind the idea is that if a block is more likely to be generated by the language model of web page, the block is more coherent with the page, i.e. more important within the whole page. The importance of different blocks is estimated by:

$$Imp(B_i, D) = \frac{\hat{p}(B_i \mid M_D)}{\sum_{j=1}^{|B|} \hat{p}(B_j \mid M_D)}$$

After that, we refined the block-level retrieval into two major components: the relevance of a block and its importance within the whole page. Another aspect of our idea is that we try to integrate user's information need to estimate block importance. Here we use query language model to estimate information need.

$$\hat{p}(t \mid M_{Q'}) = \sum_{B \in R} \hat{p}(t \mid M_B) \frac{\hat{p}(Q \mid M_B)}{\sum_{B' \in R} \hat{p}(Q \mid M_{B'})}$$

$$\hat{p}(t \mid M_{Q'}) = \sum_{D \in R} \hat{p}(t \mid M_D) \frac{Rel(D,Q)}{\sum_{D' \in R} Rel(D',Q)}$$

The new importance can be exploited in further block-level retrieval. Finally, an iterative combination can be conducted.

We also implement our approach in VSM [6] retrieval model. The topic coherence is the content similarity between block and whole page:

$$sim(D,Q) = \frac{\vec{D} \square \vec{Q}}{|\vec{D}| \times |\vec{Q}|} = \frac{\sum_{i=1}^{t} w_{i,D} \times w_{i,Q}}{\sqrt{\sum_{i=1}^{t} w_{i,D}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,Q}^2}}$$

where the vectors are composed by TF×IDF. And the iterative process is similar to that of language modeling.

4. Conclusion

In this poster, we study the issue how to combine the block importance and relevance to a given query. We propose automatic measurement of block importance and an iterative approach to combine the importance and relevance of block.

In future, we will investigate how following factors affect retrieval performance: the first is the coefficient for block importance combination; the second is the choice of relevant blocks or documents; the last is the number of iterations for iterative approaches. Besides, we will test our method on some benchmarks like TREC etc.

5. REFERENCES

[1] Cai, D., Yu, S., Wen, J.R., and Ma, W.Y. Extracting Content Structure for Web Pages based on Visual Representation. In Proceedings of the 5th APWEB, 2003.

[2] Embley, D. W., Jiang, Y., and Ng, Y.-K. Record-boundary discovery in Web documents. In Proceedings of 1999 ACM SIGMOD, 1999.

[3] Huang, S., Xue, G.-R., Yu, Y., Zhang, B., Chen, Z., and Ma, W.-Y. TSSP: A Reinforcement Algorithm to Find Related Papers. In Proceedings of the Web Intelligence, 2004.

[4] Lin, S. H., and Ho, J. M. Discovering Informative Content Blocks from Web Documents. In Proceedings of ACM SIGKDD, 2002.

[5] Ponte, J., and Croft, W. A Language Modeling Approach to Information Retrieval, in Proceedings of 21st ACM SIGIR, 1998.

[6] Salton, G., and Buckley, C. Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24(5):513-523, 1988.

[7] Song, R., Liu, H., and Wen, J.R. Learning Block Importance Models for Web Pages. In Proceedings of WWW, 2004.

[8] Xue, G.-R., Shen, D., Yang, Q., Zeng, H.-J., Chen, Z., Yu, Y., and Ma, W.-Y. IRC: An Iterative Reinforcement Categorization Algorithm for Interrelated Web Objects. In Proceedings of the ICDM, 2004.