Towards Multi-granularity Multi-facet E-Book Retrieval

Chong Huang[1], Yonghong Tian[2, 3], Zhi Zhou[2], Tiejun Huang[1, 2, 3]
[1]Graduate University, Chinese Academy of Sciences, Beijing 100039, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[3] Institute of Digital Media, Peking University, Beijing 100871, China
[1, 2, 3] {chuang, yhtian, zzhou, tjhuang}@jdl.ac.cn

ABSTRACT
Generally speaking, digital libraries have multiple granularities of semantic units: book, chapter, page, paragraph and word. However, there are two limitations of current eBook retrieval systems: (1) the granularity of retrievable units is either too big or too small, scales such as chapters, paragraphs are ignored; (2) the retrieval results should be grouped by facets to facilitate user's browsing and exploration. To overcome these limitations, we propose a multi-granularity multi-facet eBook retrieval approach.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval; H.2.8 [**Database Management**]: Database Applications – *Data Mining*

General Terms
Algorithms, Theory.

Keywords
Multi-granularity, multi-facet, e-book retrieval.

1. INTRODUCTION
Typically, a digital book (eBook) has two kinds of structures: *Hierarchy*: it has multiple granularities of semantic units - chapter, page, paragraph and word; *Hub*: each eBook is a center surrounded by its facets of properties. However, current information retrieval (IR) systems have two limitations when applied to eBook retrieval. First, the granularity of retrievable units may be either too big or too small: a book or all matched words in it. In either case, it is too tiresome for a user to scan through the whole book or search in thousands of matched but off-topic locations. Second, due to the abundance of results, grouping navigation is in need.

To overcome these limitations, we propose a Multi-granularity Multi-facet E-Book Retrieval (MMER) approach. The key to our solution is to extract facet-related information from any granularity, organize them in knowledge networks with hierarchical and radial structure, and finally provide more retrievable units and group results by facets (multi-facet navigation). Moreover, because scores of difference granularity are interrelated, we define a multi-granularity similarity metric, which can be used for multi-granularity ranking in retrieval.

2. MMER
MMER relies on three key technologies: (1) accurate extraction algorithms for both full-text and properties on any granularities; (2) effective knowledge organization model to restore relations included, especially granularity-related and facet-related information; (3) novel usage of these two kinds of information in retrieval.

Multi-granularity Information Extraction. The first issue we encounter is that only book-level metadata is assigned by librarians. Thus, we developed modules to extract information from any granularity. First, we use rule-based algorithm to separate a book into chapters or smaller granularities. Moreover, with the help of TOC files assigned by librarians, we extract the inter-granularity "belonging" relation. Second, we extract facet-related information - properties of a text using some machine learning approaches. For example, in our previous work [2], by treating a text as a semantic network, we extracted keyphrases with structural analysis on these networks and small-world model. The result is promising.

Multi-granularity Information Organization. There are three kinds of relational knowledge organization model [1]: thesauri, knowledge networks, and ontology. Thesaurus is mostly restricted in lexical analysis and ontology is suitable for more formalized and proven knowledge with complex relationships. Thus, we organize information in knowledge networks.

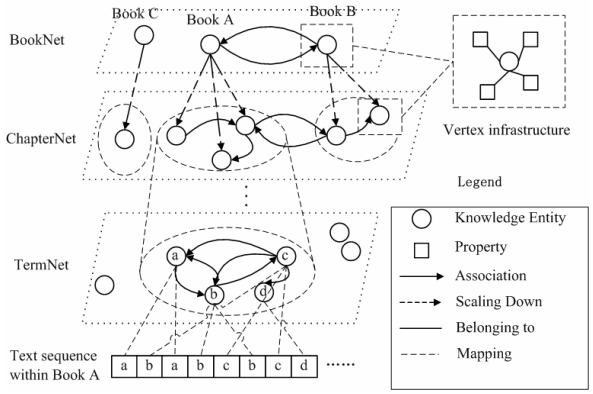


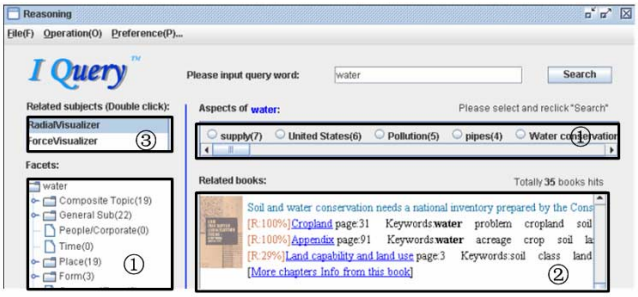
Figure 1. The structure of BookNet. The vertices in a dashed ellipse are of the same parent.

To represent hierarchical and hub-like information for eBooks, we propose a Multi-granularity Knowledge Network (MKN) model. MKN has two unique relations, namely, scaling and belonging-to. The weights of these relations are manually assigned or learn from statistical models. Unlike traditional KN, MKN provides hierarchical browsing and facet-based navigation, more accurate book similarity analysis (with relevance ranking) and more.

Two similarity functions are defined to weight the relationships in MKN. A basic similarity function measures the multi-facet similarity of two nodes in one granularity. Currently, we use a variation of cosine distance in VSM model as the basic function. Second, given difference granularities are interrelated, basic scores of related nodes on upper or lower levels are summarized with scaling weights as the final multi-granularity score of two nodes. We also include indirect relationships with a decay factor on distance, by exploring the transitivity of the similarity function. Particularly, we construct two MKNs: BookNet and SubjectNet. Books are connected in BookNet by their multi-facet similarity scores. Similarly, subjects are connected if they concur in the same book.

Copyright is held by the author/owner(s).
WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

Multi-granularity multi-facet IR. Based on knowledge in MKN, multi-granularity multi-facet IR approach returns results on both book and chapter level. It includes three key points (as in Fig.2):



Annotation: ①Facets;②Results in Book/Chapter with relevance score; ③IV Module

Figure 2. The GUI of our eBook retrieval system.

(1)Facets grouping. Because of their common values in facets (such as time, subject, etc.), eBooks are grouped by facets. Users can browse and re-search with facets on the facet tree and panel.

(2)Multi-granularity relevance analysis. A book or a chapter is ranked into a list, according to their multi-granularity similarity scores with the query. Users can access the chapters straightly.

(3)Information visualization (IV) module. Related subjects are visualized in a network style as in SubjectNet. Introducing IV into query expansion helps users reformulate his/her query.

3. A PRELIMINARY EXPERIMENT

In this experiment, we evaluate the effectiveness of MMER in the retrieval on chapter level. Except the text length and user's labor in searching, we want to clarify whether chapter returns more relevant results. We select 544 books in a wide range. Three retrieval systems are implemented: *Subject*, a matcher of the query with subjects of books; *Fulltext*, a full-text matcher; and *Chapter*, searches through extracted keyphrases from chapters as in [2].

To evaluate the effectiveness of IR system, precision and recall are usually used. However, in eBook retrieval, it is very tiresome to evaluate these two measure from all returned results and 544 books. Thus we use s@n, where s is the *relevance score* of a query and a book or a chapter (only top n in consideration). It is more accurate than binary-scored precision. Then we carry out a double-blind user scoring. For practical limitations, we have 6 users, and they select totally 15 queries. Users score a result with 2, 4, 6, 8 or 10 for relevance but no regard with the length of the result. We concentrate on three measures: ***Micro average s@10***, the average score of the top ten results for each query; ***Number of results (NR) @10***, the number results of a query, playing a similar role as recall; ***Macro average s@n***, a relevance score inter-query at a certain ranking position n.

Result 1: micro measures. Results indicate that chapter level has highest s@10 and NR@10 in most cases (9/15 and 10/15 respectively). In the table below, s(c) and s(b) stand for Mic s@10 in chapters and books returned by *Chapter*. The figures in bold are the top values in the row. Queries with an asteroid are duplicated.

Regarding accuracy, theoretically, if assigned subjects are accurate and representative, *Subject* should have the highest scores. However, s(c) of *Chapter* outperforms others in most occasions (9/15). Reasons could be: (1) Librarians and the user hold different judgment on the topic; (2) a keyword usually has different meanings in different contexts; (3) one highly-related book usually possesses several highly-related chapters.

Second, *Chapter* has a comparative NR as *Fulltext*, which is significantly higher than *Subject*. Based on empirical observations, *Fulltext* should have the highest NR. Note that some books are returned only by *Chapter* (low s(b) but high s(c)), since some highly related chapters are in seemingly unrelated books. Together with accuracy, *Chapter* returns more relevant results.

Table 1. Micro s@10 and NR@10 for each query (5 out of 15).

Query	Fulltext		Chapter			Subject	
	s	NR	s(c)	s(b)	NR	s	NR
control	3.0 ± 1.9	10	**7.2** ± 2.7	5.6 ± 2.3	10	7 ± 2	10
education*	6.3 ± 2.7	10	**8.2** ± 1.8	6.3 ± 3.0	10	8 ± 2	4
health	5.2 ± 3.3	10	5.8 ± 3.7	**6.4** ± 3.6	10	2	1
depression	5.0 ± 1.4	10	**8.0** ± 2.0	5.3 ± 1.2	3	N/A	0
sculpture	7.0 ± 3.3	10	**10.0** ± 0	**10.0** ± 0	10	**10**	1

Result 2: macro measures. In the figure below, s(c) outperforms others in all s@n ($n \le 10$). Also, s(c) has a lowest variance. The trends of these lines differ when the n is growing bigger: *Fulltext* going down drastically; the two lines of *Chapter* touch down a little and stay in a relatively steady way; *Subject* has fluctuations with an increasing trend.

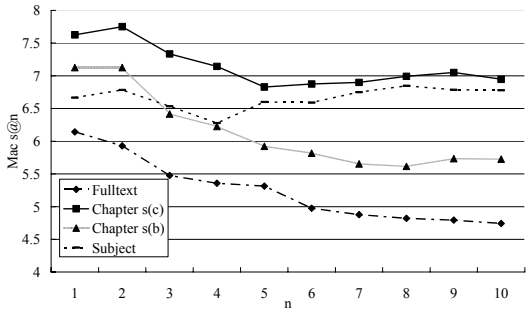


Figure 3. Macro average s@n($n \le 10$) for all three systems.

Result 3: user variance. Thanks to the existence of common queries (3/15) of different user, we can further study the variance of their relevance judgment. We use a vector to represent the scores of a query. Then the variance between two users is the cosine value of their score vectors. The result is surprising. The cosines values are all very near 1, which means they made similar judgments. Therefore, the scores in result 1 and result 2 are somehow objective and trustable.

4. ACKNOWLEDGMENTS

Our thanks to NSFC (No. 60605020) and The 863 Project of China (No. 2006AA01Z320 and No.2006AA010105) for funds.

5. REFERENCES

[1] Hodge, G. Systems of knowledge organization for digital libraries. Digital Library Federation, USA, 2000.

[2] Huang, C., Tian, Y., Zhou, Z., Ling, C., Huang, T. Keyphrase extraction using semantic networks structure analysis. In *Proc. of the ICDM'06*, pp. 275-284, Hongkong, 2006.