Towards Multi-granularity Multi-facet E-Book Retrieval in China-US Million Book Digital Library

Yonghong Tian
The Institute of Digital Media
School of EE & CS
Peking University
Beijing 100871, China
yhtian@pku.edu.cn

Tiejun Huang*
The Institute of Digital Media
School of EE & CS
Peking University
Beijing 100871, China
tjhuang@pku.edu.cn

Wen Gao
The Institute of Digital Media
School of EE & CS
Peking University
Beijing 100871, China
wgao@pku.edu.cn

ABSTRACT

There are more than one million digitalized books (i.e. e-books) so far in China-US Million Book Digital Library Project (MBP for short). It is thus important to design effective and powerful tools that enable users to easily search the required information and appropriately access knowledge in the digital library. Towards this end, currently most digital libraries simply use the traditional metadata-based or fulltext-based retrieval technologies on the e-book collection. However, there are at least two limitations of such e-book retrieval systems. (1) The granularity of retrieval results is either too big or too small, and consequently the middle granularities such as chapters and paragraphs are ignored in the traditional e-book retrieval systems. (2) The mass of retrieval results are usually ill-organized so that users often need to pay more efforts to obtain the required items. Therefore, with the many complex data in MBP, new search models and algorithms need to be developed that can take advantage of the particularities of e-books, access them appropriately, and provide results efficiently. To tackle this challenge, this paper introduces our multi-granularity and multi-aspect e-book retrieval approach for MBP. Firstly, a Multi-granularity Multi-facet Knowledge Network (MMKN) model is proposed to represent content from different granularities (e.g., books, chapters, pages, paragraphs and words) and different facets (e.g., time, space, etc.) to support retrieval of relevant items from an e-book collection. Then we implement a novel e-book retrieval system, called *IQuery*, to extract facet-related information from e-books at several granularities and then support multi-granularity e-book retrieval with more retrievable units and multi-facet navigation. Experiments were conducted to validate the efficiency and effectiveness of the proposed MMKN model, as well as the performance of IQuery. The results are encouraging, demonstrating that IQuery can provide powerful capabilities for e-book retrieval in MBP.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Digital Libraries –*Systems issues*; H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval; H.2.8 [**Database Management**]: Database Applications – *Data Mining*

General Terms

Algorithms, Theory.

ICUDL 2007, Nov 2--4, 2007, Pittsburgh, PA.

*Prof. Tiejun Huang is the corresponding author.

Keywords

Multi-granularity, multi-facet, knowledge network model, e-book retrieval.

1. INTRODUCTION

There are more than one million digitalized books (i.e. e-books) so far in China-US Million Book Digital Library Project (MBP for short). At this point, one might easily expect that in on long-term future when the full potential of MBP is realized, any citizen will be able to access all human knowledge saved in the library. However, every one in real world can only read a very small portion of books (averagely less than ten thousand books) in a digital library throughout his life [2]. Thus the digital library needs effective and efficient knowledge organization and retrieval tools to realize the mapping from one million and even more books in the library to ten thousand books for each reader. In particular, an important but urgent task for MBP is to design effective and powerful tools that enable users to easily search the required information from such a large e-book collection.

To tackle this task, currently most digital libraries simply use the traditional metadata-based or fulltext-based retrieval technologies on the e-book collection. However, there are at least two limitations of such e-book retrieval systems. First, the granularity of retrieval results is either too big or too small. The purpose of e-book retrieval is to locate the items of interest. In metadata-based or fulltext-based retrieval systems, however, the situation goes into two extremes: to return a whole book or all matched words in it. In the former situation, it is too tiresome for a user to skim through the whole book to locate the required items. On the other hand, it is also too laborious for a user to search in thousands of matched word locations, most of which are usually off-topic. Second, due to the overwhelming abundance of retrieval results, some kind of grouping navigation is in need. For example, information items within e-books can be spilt into different facets such as time, space, etc., which can be used to group the retrieval results.

Different with a web page, an e-book often has complex semantic structure. For example, each e-book has multiple granularities of semantic units — chapters, pages, paragraphs and words. Moreover, each e-book is a center surrounded by different facets of properties. Without loss of generality, the two kinds of e-book structure are referred to as *hierarchy* and *hubris* respectively. Therefore, with so many complex data in MBP, new search models and algorithms need to be developed that can take advantage of the particularities of e-books, access them appropriately, and provide results efficiently.

A possible solution is to integrate the knowledge organization system (KOS) into search models and structures. Generally speaking, the KOS has a single purpose "to organize content to support retrieval of relevant items from a digital library collection"[1]. As a typical instance of KOSs, knowledge networks (KNs) can be used to represent complex relationships between objects, e.g., the equivalence and associative relationships among terms or concepts. As mentioned above, each e-book has multiple granularities of semantic units. Thus it is necessary to extend KNs to represent complex relationships between objects at different granularities. Towards this end, we propose a Multi-granularity Multi-facet Knowledge Network (MMKN) model to represent content from different granularities (e.g., books, chapters, pages, paragraphs and words) and different facets (e.g., time, space, etc.) so as to support retrieval of relevant items from an e-book collection. Using this model, we implement a novel e-book retrieval system, called *IQuery*, to extract facet-related information from e-books at several granularities and then support multi-granularity e-book retrieval with more retrievable units and multi-facet navigation.

It should be noted that multi-granularity schemes have been studied for years in image processing [13, 14] and database management [15], even in digital library [12], but little attention has been paid on their application to knowledge organization and e-book retrieval. Multi-faceted approach has also been applied to visual information analysis [7] and OAI-PHM [5]. Dakka and Ipirotis [6] propose an automatic way of constructing a multi-faceted browser of annotated images, program schedules, and web pages. Recently, Google is also experimenting with new features aimed at improving the search results on a timeline or map view[1]. However, to the best of our knowledge, IQuery is the first system for e-book retrieval by integrating multi-granularity and multi-aspect knowledge modeling methods.

The paper is organized as follows: We present our MMKN model in Section 2 and the corresponding building algorithms in Section 3. Section 4 describes the IQuery system. Experiments are described in Section 5. Finally, Section 6 concludes this paper.

2. MULTI-GRANULARITY MULTI-FACET KNOWLEDGE NETWORK MODEL

In this section, we present the Multi-granularity Multi-facet Knowledge Network (MMKN) model. To begin with, we first clarify several concepts.

Knowledge entity. A knowledge entity is a visible or invisible carrier of a certain kind of information or knowledge, such as a book in BookNet (described later).

Association. An association is a certain kind of relationship between knowledge entities at the same granularity.

Scaling. Scaling is a concept for modeling hierarchical relationship. As the scrolling up and scrolling down in data warehousing, knowledge entities in different granularities might have scaling up/down relationship. For example, if a chapter is a knowledge entity of interest, then the book is the entity scaling up from it, and the paragraphs are entities scaling down from it.

Traditionally, the complex relationships among knowledge entities can be modeled as a knowledge network (KN). Furthermore, to simultaneously model the hierarchy and hubris structures in an e-book collection, we thus propose the MMKN model to represent the complex relationships among knowledge entities from different granularities (e.g., books, chapters, pages, paragraphs and words) and different facets (e.g., time, space, etc.). An example of the MMKN model is shown in Fig. 1.

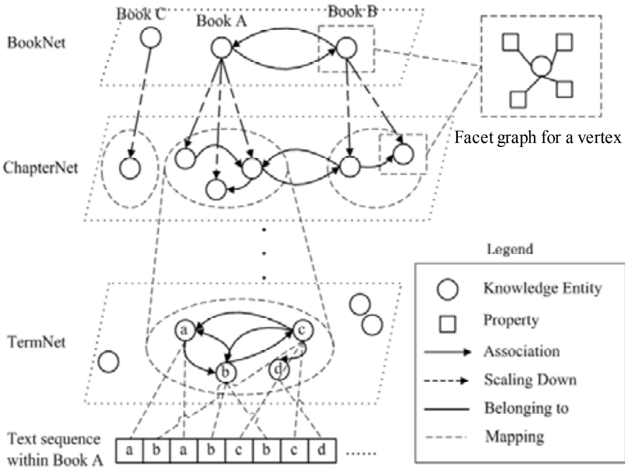


Figure 1. An example of MMKN model. The vertices in a dashed ellipse are of the same parent.

Thus we define the MMKN model formally as follows:

Vertex. Each vertex in the network represents a distinct knowledge entity. We use v_i to denote the i^{th} vertex. In MMKN, a vertex has scaling relations, pointing to its parents or children (such as components) in hierarchy.

Property. In general, each knowledge entity may possess several dimensions of *properties*. Moreover, each property might have a hierarchical structure. Thus v_i could be represented as a function of p_1, p_2 ,…, and p_n where p_1 denotes the first property of v_i. As shown in Fig 1, the properties for each vertex can be represented as a facet graph.

Edge. There are several kinds of association between two vertices, such as co-occurrence, semantic or syntactic relation, or sharing common properties. If an association exists, an *edge* links them together. Let e_{ij} denote the edge from v_i to v_j. Every edge has a *weight* w_{ij}, indicating the intensity of relationship. Most edges in MMKN are directed.

Graph (KN in one granularity). A *graph* consists of a set of vertices and a set of edges.

Distance (Shortest Path). Distance between two vertices is the length of shortest path from one to another. If edges are unweighted, a *distance* counts for number of edges that the *shortest path* passes. Here we let d_{ij} denote distance between v_i and v_j.

[1] http://www.google.com/experimental/

Mapping. There are two kinds of mapping: inter-granularity mapping by breaking a vertex into sub-vertices and vice versa; vertex-property mapping by linking a vertex to its properties.

Due to the hierarchy structure of e-books, the MMKN model consists of several layers of KNs, respectively denoted as BookNet, ChapterNet, ParagraphNet to TermNet (As shown in Fig. 1). It should be noted that for different KNs, the knowledge entity (i.e., vertex) and the association (i.e., edge) may have different meanings. For example, a vertex in BookNet represents a book, a vertex in ChapterNet represents a chapter, …, while a vertex in TermNet represents a term or a keyword. Clearly, the scaling relationship exists between two adjacent KNs (e.g., between BookNet and ChapterNet). For simplicity, the MMKN model in this paper is only restricted within three layers, i.e., BookNet, ChapterNet, and TermNet.

3. BUILDING THE MMKN MODEL

Given the definition of the MMKN model above, it is important to build an MMKN from an e-book collection. In general, building an MMKN includes three steps: (1) determining knowledge entities and their properties for different KNs; (2) establishing association between knowledge entities; (3) scoring and normalizing the weight of each association. Among them, the key issue is how to score the associations between different knowledge entities.

3.1 Scoring the associations

This paper uses similarity functions to score the association between two knowledge entities and then assign the weight of the edge among them. In general, the overall similarity between two knowledge entities is affected by three factors: the facets of two entities, the parent and children entities in hierarchy, and directness of the association. Accordingly, we develop three similarity functions.

3.1.1 Multi-faceted Similarity

Given two neighboring vertices v_i and v_j in the same granularity, we first consider how to measure the similarity between them using only multi-faceted information. Let $Sim_{\mathrm{F}}(v_i,v_j)$ denote the multi-faceted similarity function, and $Sim_k(v_i,v_j)$ the normalized similarity score between v_i and v_j on property p_k. Here $Sim_k(v_i,v_j)$ is referred to as the *basic similarity function*. Basically, there are two ways to calculate $Sim_k(v_i,v_j)$, i.e., value matching and Vector Space Model (VSM).

Value matching is an intuitive way (hit-and-gain) of scoring by matching the values of the property and its sub-properties. Simply speaking, once there is a hit in the matching, the similarity gains one point. In general, there are two levels of property match: *topical match* and *full match*. Topical match denotes the match among only the topical terms of these two properties, while full match denotes the match of all sub-properties, including the topical terms and other sub-properties (e.g., time). Empirically, similarity score between v_i and v_j on property p_k can be calculated as follows,

$$Sim_k(v_i,v_j) = \lambda_k \Delta_k(v_i,v_j)\left[1+\frac{\sum_{l=1}^{n}\Delta_l(v_i,v_j,p_k)}{n+\beta}\right], \tag{1}$$

where $\Delta_k(v_i,v_j)=\begin{cases}1, & v_i \text{ and } v_j \text{ have the same value on } p_k;\\ 0, & \textit{otherwise}\end{cases}$ denotes the number of topical matches between v_i and v_j on property p_k, and $\Delta_l(v_i,v_j,p_k)$ denotes the number of matches between v_i and v_j on the l^{th} sub-property of p_k. λ_k is a normalization factor so that $0 \le Sim_k(v_i,v_j) \le 1$ and $\sum_{\forall v_t,\exists e_{it}} Sim_k(v_i,v_t)=1$; while β is a constant with $\beta > 1$, ensuring the similarity score calculated by *k* topical matches (each of which only uses one sub-property) is larger than by one full match using all *k* sub-properties. Generally speaking, value matching is a process of hierarchical matching. As pointed out by Leung and Chen [9], the advantages of hierarchical matching are: 1) a structural comparison is made possible by matching property hierarchies; and 2) the matching process can speed up.

On the other hand, VSM represents each vertex v_i as a vector $\mathbf{v}_i$ of property values. Thus the similarity between two vertices can be calculated by the cosine value of two vectors,

$$Sim_k(v_i,v_j) = Cos(\mathbf{v}_i,\mathbf{v}_j) = \frac{\mathbf{v}_i \bullet \mathbf{v}_j}{\|\mathbf{v}_i\| \bullet \|\mathbf{v}_j\|}. \tag{2}$$

In a broad sense, value matching can be regarded as a special case of VSM where the value of each property is binary.

Then the final similarity score $Sim_{\mathrm{F}}(v_i,v_j)$ is determined by linearly combining the overall similarity scores for each property space. That is,

$$Sim_{\mathrm{F}}(v_i,v_j) = \lambda \sum_k \alpha_k Sim_k(v_i,v_j), \tag{3}$$

where α_k is a weight of property p_k for v_i. Note that $Sim_k(v_i,v_j)$ and $Sim_k(v_j,v_i)$ are often unequal due to different normalization factor λ. Note that the weight α_k is used to measure the importance of property p_k for the knowledge entity v_i. For example, the *title* property is more important for an e-book than the *publisher* property.

3.1.2 Multi-granularity Similarity

As for multi-granularity similarity between neighboring vertices v_i and v_j, the affects of inter-granularity also should be taken into account. Without risk of confusion, we use $S_+(v_i)$ denote the parent vertex set of v_i in hierarchy, and $S_-(v_i)$ the children vertex set of v_i in hierarchy. Thus the multi-granularity similarity between v_i and v_j, denoted by $Sim_{\mathrm{G}}(v_i,v_j)$, can be calculated by

$$Sim_{\mathrm{G}}(v_i,v_j) = \sum_{\substack{((\bar{v}_s\in S_+(v_i))\cap(\bar{v}_t\in S_+(v_j)))\cup\\((\bar{v}_s\in S_-(v_i))\cap(\bar{v}_t\in S_-(v_j)))}} \sigma_{i,j} Sim_{\mathrm{F}}(\bar{v}_s,\bar{v}_t), \tag{4}$$

where $Sim_{\mathrm{F}}(\bar{v}_s,\bar{v}_t)$ is the multi-faceted similarity score between $\bar{v}_s$ and $\bar{v}_t$ (here $(\bar{v}_s \in S_+(v_i))\cap(\bar{v}_t \in S_+(v_j))$ or $(\bar{v}_s \in S_-(v_i))\cap(\bar{v}_t \in S_-(v_j))$), $\sigma_{i,j}$ is a parameter to control how $\bar{v}_s$ is related to v_i and $\bar{v}_t$ to v_j. Intuitively speaking, if there is strong semantic dependency between $\bar{v}_s$ and v_i (or $\bar{v}_t$ and v_j), then $\sigma_{i,j}$ will be assign a larger value. Thus in this paper, $\sigma_{i,j}$ can be estimated approximately by

$$\sigma_{i,j} = P(\bar{v}_s \mid v_i) \bullet P(\bar{v}_t \mid v_j) \ , \qquad (5)$$

where $P(\bar{v}_s \mid v_i)$ or $P(\bar{v}_t \mid v_j)$ can be estimated by the classical probability estimation algorithms such as Bayesian Networks.

At this point, we have two similarity scores between v_i and v_j, i.e., $Sim_{\mathrm{F}}(v_i, v_j)$ and $Sim_{\mathrm{G}}(v_i, v_j)$. Clearly, the combination of multi-faceted similarity score and multi-granularity similarity score may be used to more accurately measure the association between v_i and v_j. The simplest method to calculate the overall similarity score $Sim(v_i, v_j)$ is the convex combination of $Sim_{\mathrm{F}}(v_i, v_j)$ and $Sim_{\mathrm{G}}(v_i, v_j)$, i.e.,

$$Sim(v_i, v_j) = \tau Sim_{\mathrm{F}}(v_i, v_j) + (1-\tau) Sim_{\mathrm{G}}(v_i, v_j) \ , \qquad (6)$$

where τ is the combination weight by $\tau \in (0.5, 1]$. $\tau = 1$ means the multi-granularity similarity is not taken into account. A more complex combination method is the iterative similarity propagation [10]. However, this method also suffers from much higher complexity due to the iterative computation. We thus do not intend to apply it in this paper.

3.1.3 *Multi-step Similarity*

We have now defined the similarity functions for straight association between two entities. However, the similarity is transitive. For example, if book A and book B, or book B and book C have some common properties, it can be safely deduced that book A and book C are potentially related. Without loss of generality, we refer to this as *indirect association*, and accordingly the similarity derived from indirect association as *multi-step similarity*. It is necessary to take multi-step similarity into account when the graph (i.e., KN in one granularity) is sparse.

Several multi-step similarity functions have been suggested over the network data [3], such as SimRank, Companion, Jaccard coefficient, etc. This paper uses SimRank [4] to measure the multi-step similarity among knowledge entities. For simplicity, we use $Sim^{(l)}(v_i, v_i)$ to denote the multi-step similarity with the step length *l*. Thus according to [4], the recursive SimRank iteration propagates similarity scores with a constant decay factor $c \in (0,1)$ for vertices $v_i \neq v_j$,

$$Sim^{(l+1)}(v_i, v_j) = \frac{c}{|I(v_i)||I(v_j)|} \sum_{v' \in I(v_i)} \sum_{v'' \in I(v_j)} Sim^{(l)}(v', v'') \ , \qquad (7)$$

where $I(x)$ denotes the set of vertices linking to *x*. It should be noted that if $v_i = v_j$, then $Sim^{(l+1)}(v_i, v_j) = 1$; and if $I(v_i)$ or $I(v_j)$ is empty, then $Sim^{(l+1)}(v_i, v_j) = 0$. The SimRank iteration starts with $Sim^{(0)}(v_i, v_j) = 1$ for $v_i = v_j$ and $Sim^{(0)}(v_i, v_j) = 0$ otherwise. In practice, $Sim^{(1)}(v_i, v_j)$ is calculated by Eq. (6) if $v_i \in I(v_j)$ or $v_j \in I(v_i)$. In [4], the final SimRank score is defined as the limit $\lim_{l \to \infty} Sim^{(l)}(v_i, v_j)$, but in our application the final multi-step score is controlled by the parameter *l*.

A key issue here is how to determine the decay factor *c*. The decay factor is introduced in SimRank to ensure the weight of association decays as the depth increases. There are several common decay functions (DFs): linear DF (LDF), polynomial DF (PDF), and exponential DF (EDF). They differ in the speed of decay. These functions are formulated as follows,

$$LDF(l) = \frac{1}{l} \ ,$$

$$PDF(l) = \frac{1}{poly(l)} \ ,$$

$$EDF(l) = e^{-l} \ ,$$

where $poly(l)$ is a polynomial function depending on *l*.

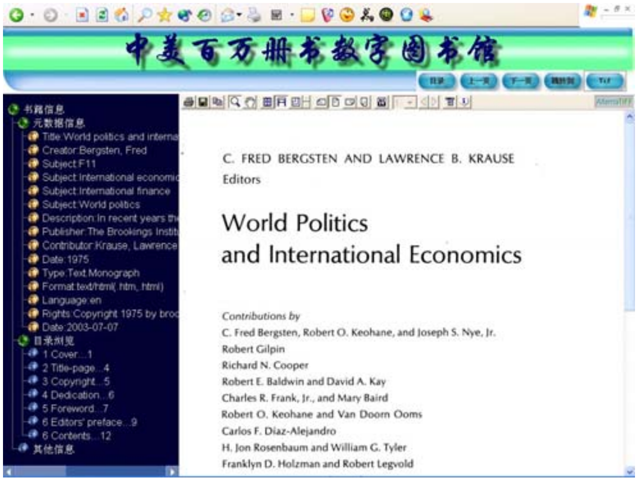
3.2 Exploiting Multi-level Information

Given the similarity functions, we need to determine how to get the properties of knowledge entities. In MBP, we can use three levels of information: manually-labeled metadata, automatically extracted keyword by the key-phrase extraction system [8], and full text.

In MBP, the metadata of each e-book is available, which is produced in the book digitalization process according to the USMARC21 entries of Library of Congress (LC) or OCLC. Some semantic-relevant fields are selected out as properties of knowledge entities, such as field 650 and its sub-fields (See Table 1). Moreover, for each e-book, there is a file “TOC.xml” to depict chapter information, which can be used to split each e-book into several chapters. An example of TOC.xml file is shown in Fig 2.

Table 1. The sub-fields of field 650 of USMARC21

$a	$v	$x	$y	$z	$2
Topical term	Form	General	Chronological	Geographic	Source of Term



中美百万册书数字图书馆

书籍信息
元数据信息
Title:World politics and interna
Creator:Bergsten, Fred
Subject:F11
Subject:International economic
Subject:International finance
Subject:World politics
Description:In recent years th
Publisher:The Brookings Instit
Contributor:Krause, Lawrence
Date:1975
Type:Text.Monograph
Format:text/html(.htm,.html)
Language:en
Rights:Copyright 1975 by broo
Date:2003-07-07
目录浏览
1 Cover...1
2 Title-page...4
3 Copyright...5
4 Dedication...6
5 Foreword...7
6 Editors' preface...9
6 Contents...12
其他信息

C. FRED BERGSTEN AND LAWRENCE B. KRAUSE
Editors

World Politics
and International Economics

Contributions by
C. Fred Bergsten, Robert O. Keohane, and Joseph S. Nye, Jr.
Robert Gilpin
Richard N. Cooper
Robert E. Baldwin and David A. Kay
Charles R. Frank, Jr., and Mary Baird
Robert O. Keohane and Van Doorn Ooms
Carlos F. Diaz-Alejandro
H. Jon Rosenbaum and William G. Tyler
Franklyn D. Holzman and Robert Legvold

Figure 2. An example of an e-book with TOC.xml.

However, there are some practical limitations for manually-labeled metadata: (1) The quality of metadata is not always satisfactory. In MBP, some e-books do not assigned to the corresponding MARC entries due to human factors in the book digitalization phase or the absence of the corresponding MARC entries in LC or OCLC. (2) In some cases, several important fields or sub-fields are missing or incomplete, even the MARC entries of these e-books are available. (3) Metadata is not provided for other granularities such as chapters or paragraphs except a whole e-book.

In previous work [8], we developed an effective and efficient algorithm to extract keywords from texts of any granularity. This algorithm treats each document as a semantic network that holds syntactic relation in edges and frequency information in nodes, and then exploits the network structure analysis models to extract key phrases. Experiments demonstrate the proposed algorithm averagely improves 50% in effectiveness and 30% in efficiency in unsupervised tasks and performs comparatively with supervised extractors. Therefore, this key-phrase extractor algorithm is used to extract keywords from texts of an e-book, a chapter or a paragraph. Naturally, each vertex is treated as a vector of key phases and VSM is used as the basic similarity function in this case.

3.3 Two Examples of Building Process

In the following, we use the top BookNet and the bottom TermNet in MMKN as two examples to illustrate the building process.

3.3.1 Building BookNet

In MMKN, BookNet is introduced to quantitatively represent the association between e-books. Thus the evaluation criterion is how well BookNet symbolizes association between e-books. As defined in Section 2, a vertex represents an e-book in BookNet. Then the properties of each e-book can be selected from three sources: manually-labeled metadata, keywords extracted from full text or chapters. To more accurately capture the association relationship, we implement three similarity measures between e-books: metadata based similarity, VSM similarity based on keywords of chapters, and VSM similarity based on keywords in full text.

As mentioned above, the metadata of each e-book is available from the corresponding MARC record. For semantic analysis of e-books, only subject fields (i.e., 6xx fields) and their sub-fields in MARC entries are used in BookNet building process. As shown in Table 1, different sub-fields of 6xx fields represent different facets about e-books. In this case, value matching method is used in the basic similarity function. Table 2 shows several examples of similarity scoring results by using Eq. (1). In the first row, the properties of two vertices are all matched, thus its scoring weight is 1.0; In the second row, only one property "Political parties" and its exclusive sub-property "Great Britain" of two vertices are matched, thus its scoring weight is smaller; While in the last row, the two vertices only share one property "Political parties" but the Geographic sub-property of the right book is not equal to "Great Britain", thus its scoring weight is smallest.

Table 2. Examples of similarity scoring in BookNet

Vertex 1 (book ID)	Vertex 2 (book ID)	weight	matched subject
31014152	31010101	1.0	1.Elections 2.Political parties
31014152	31009882	0.6666667	Political parties (Great Britain)
31014152	31009870	0.5	Political parties

On the other hand, the semantics of a keyword is dependent on its context, say, whether the keyword occurs in the whole book or only a chapter. Feature weights can be used to measure the dependency. This paper uses two feature weighting schema [8]: TFIDF and SW (score function based on Small-World Phenomenon). In the SW based scoring scheme, each document is treated as a semantic network that holds both syntactic and statistical information, and the score function $S(w_i)$ captures the centrality of word w_i in the context and the role it plays in the compactness of the network. Clearly, keywords are extracted respectively from the whole e-book or one by one from chapters.

Finally, the BookNet can be built by combining different similarity scores.

3.3.2 Building TermNet

TermNet is the bottom layer in MMKN. The knowledge entities are terms or concepts other than e-books. Thus the property selection and similarity scoring are slightly different from BookNet.

In the case that terms are from the labeled metadata (i.e. MARC entries), not only the sub-subjects but also the occurrences of terms in e-books can be treated as the properties of vertices. That is, if two terms co-occur in the same e-book, it can be deduced that they may have some semantic association. Multi-step similarity is used to measure the association between two terms, where the control parameter l is set to 2 for computational simplicity. In this case, it is easy to choose an appropriate decay factor since there is no significant difference between two decay factors PDF and EDF. Figure 3 shown an example of similarity scoring for term association by using metadata,

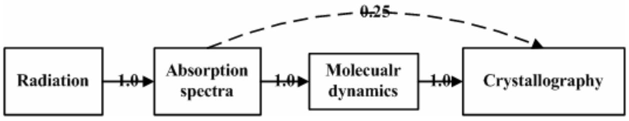


Figure 3. Similarity scoring for term association by using metadata. Solid lines are straight association, and dashed lines are indirect association.

In the case of automatically extracted keywords, keywords can be approximately treated as terms in our application. Each keyword can be viewed as a vector of documents. Thus the similarity between two keywords can be easily calculated by Eq. (2). This similarity can also be referred to as semantic proximity in [4].

Similarly, the TermNet can be built by combining different similarity scores.

4. THE IQUERY SYSTEM

As a powerful knowledge network model to represent content from different granularities and different facets, MMKN can be applied to a wide range of applications. By exploiting BookNet and TermNet, MMKN can be used to knowledge-based e-book browsing and navigation. In our previous work, KnowMap [10] is such a hierarchical e-book browsing system on the basis of MMKN model. In this section, we describe a novel e-book retrieval system, called *IQuery*, also based on MMKN model.

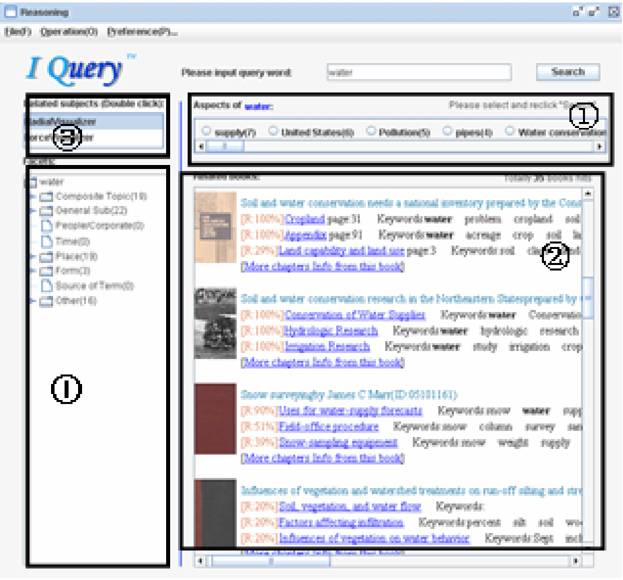
On the top of the MMKN framework, IQuery system extracts facet-related information from e-books at several granularities and then supports multi-granularity e-book retrieval with more retrievable units and multi-facet navigation. Given a query, IQuery first searches the submitted keyword through TermNet. If it hits, facets (groups of sub-subjects) in TermNet are returned. Meanwhile, the system also searches the relevant e-books or chapters from MMKN. Finally, the system returns different granularities of retrieval results and displays them in the visualization way. Compared with the traditional e-book retrieval system, IQuery has three key modules.

(1) **Facet grouping module**: According to their topical properties and sub-properties in the BookNet, retrieval results are grouped into different facets. Consequently, users can browse through facet navigator and further refine their inquiries according to the provided facets.

(2) **Multi-granularity relevance analysis module**: With the multi-granularity information available in MMKN, e-books and chapters are ranked and re-ordered according to their multi-granularity similarity scores with the given query. Users thus can access the chapters directly.

(3) **Information visualization module**: This module is used to visualize the semantic structure of TermNet with the given inquiry keyword as the network center, consequently facilitating users to refine their inquiries to relevant topics.

Figure 4 shows the main interface of IQuery.



Soil and water conservation needs a national inventory prepared by the Cons
[R:100%]Cropland page 31 Keywords:water problem cropland soil
[R:100%]Appendix page 91 Keywords:water acreage crop soil la
[R:20%]Land capability and land use page 3 Keywords:soil
[More chapters Info from this book]

Soil and water conservation research in the Northeastern Statesprepared by
[R:100%]Conservation of Water Supplies Keywords:water Conservatio
[R:100%]Hydrologic Research Keywords:water hydrologic research
[R:100%]Irrigation Research Keywords:water study irrigation crop
[More chapters Info from this book]

Snow surveyingby James C Marr(ID:05101161)
[R:90%]Uses for water-supply forecasts Keywords:snow water supp
[R:51%]Field-office procedure Keywords:snow column survey san
[R:39%]Snow-sampling equipment Keywords:snow weight supply
[More chapters Info from this book]

Influences of vegetation and watershed treatments on run-off silting and str
[R:20%]Soil, vegetation, and water flow Keywords:
[R:20%]Factors affecting infiltration Keywords:percent silt soil wo
[R:20%]Influences of vegetation on water behavior Keywords:Sept incl

Annotation: ①Facets, Aspects; ②Multi-scale results; ③IV Module

Figure 4. GUI of our e-book retrieval system, IQuery.

4.1 Facet Grouping

In IQuery, retrieval results are grouped into *aspects* (e.g., "water supply", "water resource in USA" and so on for the query keyword "water") according to different values of topic-related fields and subfields; then different aspects are grouped into *facets* (e.g., people, time or place) according to the meaning of these fields. As shown in Fig. 4, the aspect list can be used to query expansion, and the facet information is displayed in a navigation tree.

Different with the traditional clustering, facet grouping can be treated as a multi-view clustering technology. The same e-book or chapter can be grouped into several aspects or facets according to its different properties or different values of the same property. By using facet grouping, we can easily build a concept hierarchy. Therefore, facet grouping can provide a novel navigation way for retrieval results.

Currently, facets are defined according to fields and subfields in metadata: *composite terms* (terms including the query keyword), *time*, *place*, *general subfield* (category), *forms of reservation*, *people*, *source of topic terms* (LCSH or other), and *others* (unclassified aspects). For e-books or chapters without labeled metadata, different methods can be used to extracted aspects and facets. For example, the composite terms can be further distilled from the key-phrases that are extracted by the key-phrase extractor; the general subfields can be selected from the co-occurring keywords with the given query keywords; for some named entities, we can obtain them from the thesauri, or learn by using linguistic rules and text mining.

4.2 Query Relevance Analysis in Multi-granularity Context

In IQuery, multi-granularity information can be not only used to support e-book retrieval with different granularities of retrievable units (e.g., books, chapters or paragraphs), but also used to improve the ranking of retrieval results. Intuitively, if a user input a query "neural network", a book with several relevant chapters should be assigned to a higher rank score than another book with only one relevant chapter, even though the key-phrase "neural network" may have the same occurrence scores in the two books. In this paper, we refer to this query relevance analysis in multi-granularity context as *multi-granularity ranking*.

Book	Relevance
A	0.8
B	0.6
C	0.55

Initial relevance scores for e-books

Chapter	Relevance
A.2	0.5
B.2	0.3
B.3	0.4
C.1	0.6
C.4	0.9

Initial relevance scores for chapters
(Assume that each e-book has ten chapters)

(a) Initial ranking by using intra-granularity information

Rank	Book
1	A
2	C
3	B

(b) Re-ranking of e-books by taking into account the affects of chapter relevance scores on e-book ranking

Rank	Chapter
1	C.4
2	A.2
3	C.1
4	B.3
5	B.2

(c) Re-ranking of chapters by taking into account the affects of book relevance scores on chapter ranking

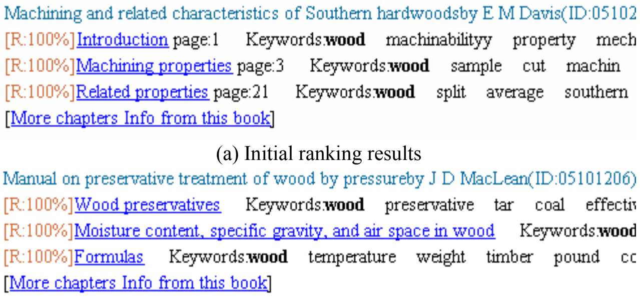
Figure 5. An example of the calculation of multi-granularity ranking.

Fig. 5 shows an example of the calculation of multi-granularity ranking. Formally, let b_i denote a relevant book in retrieval results, c_s a relevant chapter in retrieval results. Given a query q, we use $r^{(0)}(b_i \mid q)$ (or $r^{(0)}(c_s \mid q)$) to denote the initial relevance score of book b_i (or chapter c_s) regarding to q. Thus if books are treated as the retrieval units, then the relevance score $r^{(1)}(b_i \mid q)$ of multi-granularity ranking can be defined as follows:

$$r^{(1)}(b_i \mid q) = \theta r^{(0)}(b_i \mid q) + (1-\theta)\sum_{c_s \in b_i} \varpi_s r^{(0)}(c_s \mid q)\ , \qquad (8)$$

where the parameter θ is used to control the affects of chapter relevance scores on e-book ranking, ϖ_s is a weight to measure the topical dependency of the given chapter on the whole book, with $\sum \varpi_s = 1$ and $0 \le \varpi_s \le 1$. For example, $\varpi_s = 0.1$ for middle chapters, $\varpi_s = 0.05$ for introductory and summary chapters but $\varpi_s = 0.01$ for auxiliary chapters such as references and bibliography. A similar formula can be easily deduced when chapters are treated as the retrieval units.

Fig. 6 shows an example of multi-granularity ranking results, where chapters are treated as the retrieval units. We can see that in the initial ranking results, the introduction chapter is in the first rank. In most cases, the introduction chapter might be too general to satisfy the user's needs. We can see that this situation has been improved in the multi-granularity ranking results, in which chapters with more concrete content are in the first rank.



Machining and related characteristics of Southern hardwoodsby E M Davis(ID:05102

[R:100%]Introduction page:1 Keywords:**wood** machinabilityy property mech

[R:100%]Machining properties page:3 Keywords:**wood** sample cut machin

[R:100%]Related properties page:21 Keywords:**wood** split average southern

[More chapters Info from this book]

(a) Initial ranking results

Manual on preservative treatment of wood by pressureby J D MacLean(ID:05101206)

[R:100%]Wood preservatives Keywords:**wood** preservative tar coal effecti

[R:100%]Moisture content, specific gravity, and air space in wood Keywords:**wood**

[R:100%]Formulas Keywords:**wood** temperature weight timber pound cc

[More chapters Info from this book]

(b) Multi-granularity ranking results

Figure 6. An example of multi-granularity ranking results, where chapters are treated as the retrieval units.

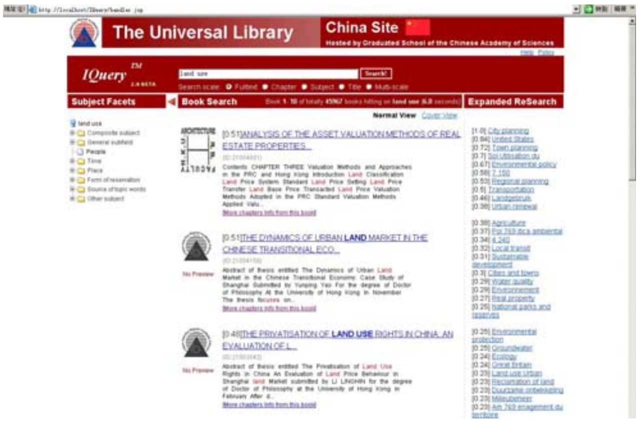
4.3 Information Visualization

Information visualization (IV) module can provide powerful capabilities that enable users to refine their inquiries, navigate the topical space related to the query, analyze the results, and change the form of the information to interact with it. However, visualizing a topic network with millions of nodes is very challenging.

Generally speaking, when a user enters a query, what he/she is most interested in is not the entire topic network, but local topical structure related to the given query. As a result, IQuery employs a centroid-driven approach to visualize the semantic structure of TermNet. That is, when a user inputs his/her query, the IV module returns a part of the topic network, with the query keywords at the center and all other vertices in a two-degree separation from the center. Developed with the open-source software prefuse (heep://prefuse.org), the IV module can effectively facilitate users to refine their inquiries to relevant topics. Fig. 8 shows an example of the centroid-driven IV panorama of TermNet, with "Petrology" as the center.

4.4 Web-based IQuery System

Fig. 7 shows the snapshot of Web-based IQuery system, which uses Tomcat as the Web service software and Lucene as the backend full-text search engine. Currently, IQuery runs on an archive with about 150,000 English e-books in MBP.



[0.51]ANALYSIS OF THE ASSET VALUATION METHODS OF REAL ESTATE PROPERTIES...

Contents CHAPTER THREE Valuation Methods and Approaches in the PRC and Hong Kong Introduction Land Classification Land Price System Standard Land Price Setting Land Price Transfer Land Base Price Transacted Land Price Valuation Methods Adopted in the PRC Standard Valuation Methods Applied Valu...

[More chapters Info from this book]

[0.51]THE DYNAMICS OF URBAN LAND MARKET IN THE CHINESE TRANSITIONAL ECO...

Abstract of thesis entitled The Dynamics of Urban Land Market in the Chinese Transitional Economy Case Study of Shanghai Submitted by Yunping Yao For the degree of Doctor of Philosophy At the University of Hong Kong In November The thesis focuses on...

[More chapters Info from this book]

[0.48]THE PRIVATISATION OF LAND USE RIGHTS IN CHINA: AN EVALUATION OF L...

Abstract of thesis entitled The Privatisation of Land Use Rights in China An Evaluation of Land Price Behaviour in Shanghai land Market submitted by LI LINGHIN for the degree of Doctor of Philosophy at the University of Hong Kong in February After d...

[More chapters Info from this book]

[1.0] City planning
[0.84] United States
[0.72] Town planning
[0.7] Soil Utilisation du
[0.67] Environmental policy
[0.58] 7.150
[0.53] Regional planning
[0.5] Transportation
[0.46] Landgebruik
[0.38] Urban renewal

[0.38] Agriculture
[0.37] Pol 769 dca ambiental
[0.34] 4.240
[0.32] Local transit
[0.31] Sustainable development
[0.3] Cities and towns
[0.29] Water quality
[0.29] Environnement
[0.27] Real property
[0.25] National parks and reserves

[0.25] Environmental protection
[0.25] Groundwater
[0.24] Ecology
[0.24] Great Britain
[0.23] Land use Urban
[0.23] Reclamation of land
[0.23] Duurzame ontwikkeling
[0.23] Milieubeheer
[0.23] Am 769 engagement du territoire

Figure 7. The snapshot of Web-based IQuery system.

5. EXPERIMENTS

Experiments were conducted to validate the efficiency and effectiveness of the proposed MMKN model, as well as the performance of IQuery.

5.1 Efficiency of MMKN Building

In this experiment, we investigate the time cost of MMKN building. We select 100, 1000, 5000, 10000 e-books from our archive, and develop an automatic builder for MMKN with JAVA. The builder runs on a PC with Intel 4 1.6GHz CPU and 1.0G memory. Results are shown in the table 3. For space limitation, here we only give the data regarding BookNet and TermNet. Note that here *l* denotes the control parameter *l* in multi-step similarity function. *l*=1 means only the direct association is taken into account, while *l*=2 means both the direct association and the indirect association with 2 steps are taken into account.

Table 3. Time cost of MMKN building.

Num of Books	Num of terms	BookNet edges	TermNet edges	BookNet Building Time	TermNet Building Time	
					l=1	*l*=2
100	176	154	874	1"	4"	5"
1000	898	12,096	12,721	2"	22"	25"
5230	3,579	77,438	50,284	7"	1'18"	1'29"
10000	16,998	265,936	110,561	18"	2'20"	2'47"

We can see that the MMKN building process is computational efficient. Moreover, the time cost of MMKN building increases much more slowly than the number of input nodes. Comparatively, the building time of TermNet is much more than BookNet. A possible reason is that they are implemented with different search algorithms. BookNet is built using Trie Tree as search data structure, while TermNet is built using SQL operations on the Term database.

5.2 Effectiveness of MMKN

To investigate the effectiveness of MMKN, experiments should be performed on its instances, such as BookNet or TermNet. However, the evaluation of the accuracy of BookNet is a goal hard to achieve. First, each book has a long but unstructured text, and it is difficult for one person to summarize the topics without skimming through the full text. It is also a very time-consuming task for such a large archive. Furthermore, the definition and criterion of topical association between books differs from one to another, even though they are all experts. Therefore, we here only validate the effectiveness of TermNet in a case study (an example of BookNet is shown in table 2).

We choose 5230 e-books from our archive, and 9516 subject field entries are available for these e-books. Then a TermNet is built on this data set. The results are shown in Fig. 8 and Fig. 9. Since the entire network is too complex and huge, we only deploy a local graph with a randomly-selected vertex, “Petrology”, as the center. In the two figures, the center node, its straight neighbors, and its two-step neighbors are marked by different padding colors. The width of each edge indicates the similarity score between two vertices.

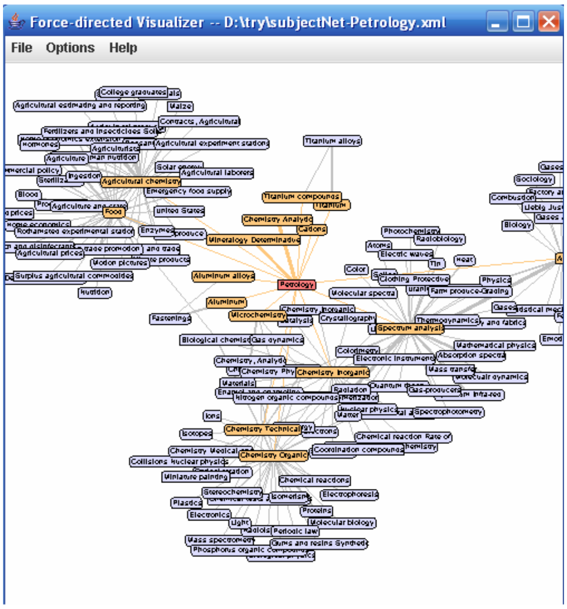

Force-directed Visualizer -- D:\try\subjectNet-Petrology.xml
File
Options
Help

Figure 8. The panorama of the centroid-driven IV result of TermNet, with “Petrology” as the center.

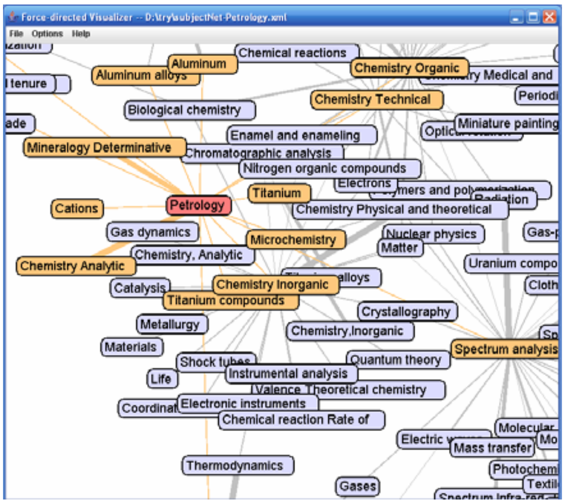


Figure 9. The enlarged local graph of the centroid-driven IV result of TermNet, with “Petrology” as the center.

From Fig. 9, we can see that “Mineralogy Determinative” and “Chemistry Analytic” have the highest similarity score with “Petrology”. It really makes sense in the real world, since the first is a theory in petrology, and the latter is one of the core technologies of petrology. Also, from the two-step neighbors, we can find some related concepts to “Petrology”, such as “Biological chemistry” (on the formation of stones) and “Enamel and enameling” (on usage of petrology).

Clearly, the abundance of concept relation relies on the size of the data set. With more e-books available, we can obtain more interesting experience in the navigation of TermNet. Some surprising but reasonable associations may spark further discoveries or new ideas.

5.3 Performance of IQuery

5.3.1 *Experimental Setup*

In this experiment, our aim is to test whether IQuery is able to improve the user’s experience in e-book retrieval task by exploiting multi-granularity and multi-facet information. Besides IQuery, two additional e-book retrieval systems are used as baselines. The first system, denoted by MDSearch, searches results for the given query on a metadata base (including all subject fields in MARC entries). The second system, denoted by FTSearch, searches results for the given query through a full-text search engine (Lucene is used in our experiments).

We select 544 books from our e-book archive, covering nearly all subjects available, from agriculture, arts, economy, engineering, history, mathematics, management and so on. Then we extract key-phrases from each chapter of these books, using key-phrase extractor proposed by [8]. In the experiment, users can access the full text, chapter information, and metadata of each book.

5.3.2 *Evaluation Measures*

To evaluate the effectiveness of information retrieval system, precision and recall are usually used. However, in e-book retrieval, it is very tiresome for one to skim through all the books to determine how many of them are related to a query. Moreover, most users won’t be patient enough to browse through all returned results. As a result, some literatures choose top *n* precision, denoted by p@*n*, as the evaluation measure. p@*n* evaluates how many results on the top *n* are relevant to the given query. However, we use a variation of p@*n* – *s*@*n*, where *s* is the *score of relevance* between a query and a book or a chapter. We argue that here relevance score is more accurate than precision that uses binary scores.

As for the criterion of relevance, we carry out a double-blind user survey. We invite users to input any query word they like, and score top 10 returned results. Note that users are not aware of the technical backgrounds of each retrieval system and which one or two are baseline(s). For practical limitations, we have 6 users to fulfill this experiment. The value scope of relevance scores is constrained in [0, 10]. As psychology studies, users are prone to choose middle score 5, so we choose discrete relevance scores including 2, 4, 6, 8, and 10. The meaning of each score is listed in Table 4. Users can make their decisions based on the metadata, table of contents, and full text of each book or chapter. As for Chapter, users are instructed to score the corresponding book of the returned chapter first and then the chapter. Otherwise, users will easily over-score the book because of the intervention of the chapters.

Table 4. Definition of optional Relevance Score and mapping between two kinds of scores.

Score	2	4	6	8	10
definition	Not relevant	A little relevant	Mediated relevant	Quite relevant	Very relevant

Then three measures are used.

Micro average s@10 (Mic s@10). Usually one page displays 10 results. Therefore, we investigate only the scores in the first page. Micro s@10 means the average score of the top 10 results for each query.

Number of result @10 (NoR@10). It indicates the number results of each query, which plays a similar role as recall. NoR is sometimes smaller than 10.

Macro average s@n (Mac s@n, $n \leq 10$). It shows the average score of the top n results for all queries.

In addition, if some users choose the same query, we can investigate scoring variance for different users.

5.3.3 Experimental Results

Result 1: micro measures. The result of micro measures is shown in Table 5. In this table, there are several acronyms: s is short for Mic s@10; s(c) and s(b) stand for Mic s@10 when returned chapters and books respectively; N is short for NoR@10. The figures in bold are the top values in the row. The queries with an asterisk are duplicated queries. When the query is "China", FTSearch does not return any results (denoted by N/A) possibly due to case recognition failure.

Table 5. Micro measures @10 for each query.

Query	**FTSearch**		**IQuery**			**MDSearch**	
	s	**N**	**s(c)**	**s(b)**	**N**	**s**	**N**
control	3.0 ± 1.9	10	**7.2** ± 2.7	5.6 ± 2.3	10	7 ± 2	10
mathematics	4.8 ± 3.4	10	N/A	N/A	0	**6** ± 3	9
beauty	4.8 ± 2.5	10	**5.2** ± 2.1	4.6 ± 1.6	10	N/A	0
education*	6.3 ± 2.7	10	**8.2** ± 1.8	6.3 ± 3.0	10	8 ± 2	4
multimedia	N/A	0	N/A	N/A	0	N/A	0
protocol	N/A	0	N/A	N/A	0	N/A	0
China*	N/A	0	6.0 ± 2.1	4.4 ± 0.8	5	**10**	1
culture	**5.8** ± 2.7	10	5.4 ± 3.3	4.2 ± 2.6	10	3 ± 1	3
war*	2.0 ± 0	10	**7.7** ± 2.5	6.0 ± 2.9	10	7 ± 3	10
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
vitamin	4.6 ± 2.5	10	**6.8** ± 2.9	5.8 ± 2.4	10	4	1
population	5.2 ± 3.6	10	**7.0** ± 3.7	6.0 ± 3.7	10	6 ± 6	2
symphony	4.4 ± 2.8	10	**8.0** ± 0	**8.0** ± 0	3	N/A	0
sculpture	7.0 ± 3.3	10	**10.0** ± 0	**10.0** ± 0	10	**10**	1

Note: s is short for Mic s@10; s(c) and s(b) stand for Mic s@10 when returned chapters and books respectively; N is short for NoR@10.

As for retrieval accuracy, MDSearch should have the highest scores, since theoretically the manually-labeled topic terms should best capture the content of books. However, the result is that s(c) of IQuery outperforms others in most cases (9/15). There might be several reasons. First, the subject fields in metadata are missing or incomplete in many books, leading to the fluctuant performance of MDSearch. Second, a word usually has different meanings in different contexts. Since MDSearch employs the word matching method without taking its context into account, the retrieval results will be certainly assigned to low relevant scores by users. Finally, one highly-relevant book usually has several highly-relevant chapters. Thus when more than one of these chapters is returned in IQuery, it will obtain a higher mic s@10. An interesting but natural finding in the experiment is that for real users, chapters seem to be a more suitable retrievable unit than books. Consequently IQuery has signification advantages in relevance scores of returned results since it can quickly locate the required items.

As for NoR, both IQuery and FTSearch are significantly higher than MDSearch. Empirically speaking, FTSearch should have the highest recall (NoR), since it employs the largest word set. Overall, IQuery returns more results with higher relevance scores.

But IQuery still has its shortcomings when given some general words such as "mathematics". The reason lies in the mismatching between stemming and de-stemming. We can see "mathematic" and "mathematical" as keywords in the chapter, but our current stemmer and de-stemmer fail to map them to the same stem. This means that there is still some room for improvement for IQuery.

Result 2: macro measures. In the Fig. 10, we can see that IQuery with s(c) outperforms others in all top@n ($n \leq 10$).The variances of Mac s@n for FTSearch, IQuery with s(c), IQuery with s(b), and MDSearch are in the scope of [2.87, 3.37], [2.33, 2.81], [2.62, 2.81], and [3.04, 3.60], respectively. Clearly, IQuery with s(c) has a relatively low variance.

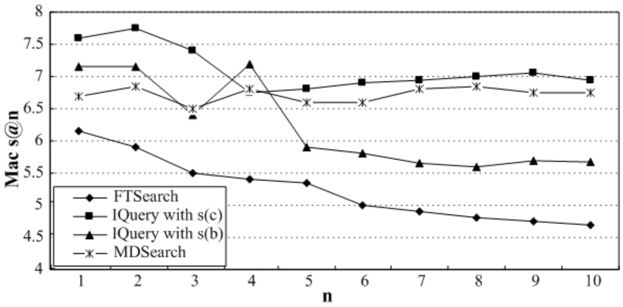


Figure 10. Macro average s@n for all three systems.

Result 3: user variance. Thanks to the existence of replicated queries, we can further study the variance of relevance scores by different users on the same query. Intuitively, if we treat the relevance scores of the 10 results for each replicated query (though sometimes it is less than 10) as a score vector, then the variance between two users can be calculated by the cosine value of these two vectors. The result is surprising. As shown in Table 6, the cosines values are all very near 1. This means that different users have the similar relevant scores for the same ranking of retrieval results. To some extent, this validates the fact that the above two experimental results have a comparatively strong generalization.

A possible reason might be that users with similar educational background tend to make similar judgments on relevance scores.

Table 6. Relevance score variance between users.

Query	FTSearch	IQuery		MDSearch
		s(c)	s(b)	
education	0.854	0.938	0.931	0.990
China	N/A	0.963	0.980	1.000
war	1.000	0.930	0.929	0.990

In summary, the experimental results are generally positive, but in some cases, the improvements are not so significant. However, we can safely conclude from these results that IQuery can provide novel and powerful capabilities for e-book retrieval in MBP.

6. CONCLUSION

As the number of digitalized e-books increases rapidly in MBP, it is crucial to design effective and powerful tools that enable users to easily search the required information from such a large e-book collection. This paper presents our multi-granularity and multi-aspect e-book retrieval approach for MBP. A novel system, called IQuery, has been implemented to extract facet-related information from e-books at several granularities and then support multi-granularity e-book retrieval with more retrievable units and multi-facet navigation. Experimental results show that IQuery can provide powerful capabilities for e-book retrieval in MBP.

In future, we will investigate how to extend IQuery on a larger e-book archive with multiple languages. Clearly, IQuery is only a start point towards developing new search models and structures that can support effective and efficient knowledge organization and retrieval in MBP by taking advantage of the particularities of e-books.

7. ACKNOWLEDGMENTS

This work is supported by China-America Digital Academic Library (CADAL) project, also supported by grants from Chinese NSF under contract No. 60605020, National Hi-Tech R&D Program (863) of China under contract No. 2006AA01Z320 and 2006AA010105, and National Key Technology R&D Program under contract No. 2006BAH02A10. We are grateful to Chong Huang, Zhi Zhou for system implementation.

8. REFERENCES

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

[2] Huang, T., Tian, Y., et al. Towards a multilingual, multimedia and multimodal digital library platform. *J. Zhejiang Univ. SCI 2005 6A*(11):1188-1192.

[3] Liben-Nowell, D. and Kleinberg, J. The link prediction problem for social networks. In *Proc. of CIKM*, pages 556-559. ACM Press, 2003.

[4] Kandola, J., Shawe-Talyor, J., & Cristianini, N. Learning semantic similarity. In *Proc. of Int'l Conf. Advances in Information Processing System (NIPS 15)*, 2002, pp. 673–680, MIT Press: Cambridge, MA, USA.

[5] Jerez, H., Liu, X., et al. The multi-faceted use of the OAI-PMH in the LANL repository. In *Proc. of JCDL'04*, Tucson, Arizona, USA, 2004.

[6] Dakka, W., Ipirotis, P.G., Wood, K.R. Automatic construction of multifaceted browsing interfaces. In *Proc. of CIKM '05*, Nov. 2005.

[7] Hetzler, E., Whitney, P. Multi-faceted insight through interoperable visual information analysis paradigms. In *Proc. of the 1998 IEEE Symposium on Information Visualization*, October 1998.

[8] Huang, C., Tian, Y., Zhou, Z., Ling, C., Huang, T. Keyphrase extraction using semantic networks structure analysis. In *Proc. of the sixth IEEE Int'l Conf. on Data Mining*, Hongkong, 2006.

[9] Leung, W.-H. and Chen, T. Hierarchical matching for retrieval of hand-drawn sketches. In *Proc. of 2003 International Conference on Multimedia and Expo (ICME '03)*, July 2003, vol.2, II- 29-32.

[10] Wang, X.J., Ma, W.Y., Xue, G.R., Li, X.. Multi-Model Similarity Propagation and its Application for Web Image Retrieval. *In Proc. of 12th ACM International Conference on Multimedia*, New York, USA, p.944-951.

[11] Ramadan, E., Tarafdar, A., Pothen, A., A hypergraph model for the yeast protein complex network. In *Proc. of Parallel and Distributed Processing Symposium'04*, 2004.

[12] Smith, T. R. A Digital Library for Geographically Referenced Materials. *IEEE Computer*, 29 (1996):54-60

[13] Malat, S., Zhong, S. Characterization of signals from Multigranularity edges. *IEEE Trans on PAMI*, 1992, 14(9):710-732.

[14] Choi, H., Baraniuk, R.G. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Tran. on Image Processing*, 10(9), 2001, pp. 1309 -1321.

[15] Zhou, S., Jones, C. B. Design and implementation of Multigranularity database. In *Proc. of the 7th International Symposium on Spatial and Temporal Database*, SSTD 2001.