Searching Fuzzy Information in Digital Library

Do Quang Vinh
Deparment of Information Technology
Hanoi University of Culture
418 La Thanh Street, Dong Da, Ha Noi, Viet Nam
Email: vinhdq@huc.edu.vn

ABSTRACT

Now, method of retrieving and collecting information has changed. It's not necessary to go out for searching and accessing large available information online via portal, provided by several information providers such as: DL, digital publishers, enterprises, organizations, individuals. Accessing information is not limited by available books or magazines in the nearest library, it can be accessed via big databases and documents distributed in over the world. It's not only texts and digital data, but also includes images, sounds/voices, geographic data, video, audio, multimedia. It enables users to go a virtual travel in the museums, historic spots and wonder of natures, to participate in the concerts and virtual performance, to watch movies and read books, to listen the lectures and music - all via DL.

KEYWORDS

digital library, model of retrieving information, method of retrieving information, retrieving fuzzy information, searching with near operator

1. DEFINITION OF DIGITAL LIBRARY

1.1 Informal Definition

Herein, we introduce informal definitions on DL.

Definition 1 (Arms W.Y.) [1]: digital library is a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network. The main idea is the managed information. Digital libraries contain diverse collections of information for use by many different users. Digital libraries range in size from tiny to huge. They can use any type of computing equipment and any suitable software. The unifying theme is that information is organized on computers and available over a network, with procedures to select the document in the collections, to organize it, to make it available to users, and to archive it.

Definition 2 (Reddy R., Wladawsky-Berger I.) [3,10]: DL is network document storage on digital documents, image, sound, science data and software that are the core of today's Internet and tomorrow's universally accessible digital repositories of all human knowledge.

Definition 3 (The Digital Library Federation) [3,10]: Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

In conclusion, Digital Library is a huge managed collection of digital information with associated services.

1.2 Formal Definition

Next, we introduce formal definition on DL:

Digital Library is a set of four (R, MC, DV, XH) in which:

- R is a storage;
- MC is index of meta-data;
- DV is a collection of services containing indexing, searching and browsing services;
- XH is a user community of digital library.

ISBN: 978-1-941968-42-0 ©2017 SDIWC

2. MODEL OF INFORMATION RETRIEVAL

Information retrieval (IR) refers to managing, storing, searching and evaluating relevant information to user's demand. [2, 4, 5, 6, 7]

Overall model of information retrieval is a pair including objects and an search mapping to some objects with an representative object for a query.

Given

$$D = \{d_1, d_2, ..., d_M\}, M \geq 2 \quad (1)$$

is a limited non-empty collection of object.

Note: case of M = 1 can be considered, but it's normal. Typical objects are representation.

Given $\Re$ as a search mapping from D in $\rho(D)$, it means that,

$$\Re : D \rightarrow \rho(D). \quad (2)$$

By combining collection of D objects and $\Re$ search mapping, we define structure of information retrieval as follows:

Definition 3.1 (structure of information retrieval):

Structure of information retrieval (SIR) is a set $2\ S = <D, \Re>$. (3)

Definition 3.1 is a general definition: it doesn't mention about distinct type of $\Re$ search mapping and object D. So, it can be obtained the different types of model IR by specifying D and $\Re$.

We introduce a consistent definition on IR models using SIR.

Definition 3.2 (Model of information retrieval - MIR):

Model of information retrieval (MIR) is a SIR $S = <D, \Re>$ with 2 following attributes:

(i) $q = \delta \Rightarrow \mu\tilde{a}_i(q, \delta) = 1\ \forall i, q, \delta$ (mapping); (4)

(ii) $\Re^i(q) = \{\delta \in D | \mu\tilde{a}_i(q, \delta) = \max \mu\tilde{a}_k(q, \delta_k)\} \cap a\alpha_i$, optional fixed i.

in which:

+ $T = \{t_1, t_2, ..., t_N\}$ is a limited collection of index term, $N \geq 1$;

+ $O = \{o_1, o_2, ..., o_U\}$ is a limited collection of object, $U \geq 2$;

+ $(D_j)_{j \in J = \{1, 2, ..., M\}}$ is a cluster group of object, $D_j \in \rho(O)$, $M \geq 2$;

+ $D = \{\delta_j | j \in J\}$ is a collection of document, in which fuzzy collection is standardized $\delta_j = \{(t_k, \mu\delta_j(t_k)) | t_k \in T, k = 1, ..., N\}, j = 1, ..., M, \mu\delta_j : T \rightarrow S \subseteq [0, 1] \subset \mathbf{R}$ is cluster representation of cluster object D_j. For example, O may include articles, each article is a cluster and each of fuzzy representation of cluster is a document. In this case, if fuzzy collection is exact collection, the document is unique to presentation of classic binary vector. Other example, cluster can be a collection of related articles, in which representation of cluster or document is a fuzzy representation for one of articles or false articles (or it was not an article in cluster but it has good description on the whole content of cluster). Therefore, it is a special case of traditional model of IR cluster.

+ $A = \{\tilde{a}_1, ..., \tilde{a}_C\}$ is a limited standard collection, $C \geq 1$, in which $\tilde{a}_i = \{((q, \delta_j), \mu\tilde{a}_i(q, \delta_j)) | \delta_j \in D, j = 1, ..., M\}, i = 1, ..., C$ is a fuzzy standardized relation, $\mu\tilde{a}_i : D \times D \rightarrow [0, 1] \subset \mathbf{R}$, $q \in D$ optional fixed. In addition, classic IR has splitting attribute (bipolar) in which there are 2 clear standards:

(i) existence and non-existence;

(ii) searching depends on (i).

We assume that it has more than 2 standards (or relevant, irrelevant, undeterminable) with different levels. So, we have to accept the standard of fuzzy relation.

+ $a\alpha_i = \{\delta \in D | \mu\tilde{a}_i(q, \delta) > \alpha_i\}, i = 1, ...,$ C is a α_i-strong standard section $\tilde{a}_i$, $\alpha_i \geq 0$, $q \in D$ optional fixed;

+ $\Re : D \rightarrow \rho(D)$ is a search mapping. In general view, searching is link of subset document and a query if they are linked - according to a strong enough selection standard. So that, we have to consider that query which is a document and search, is defined to use α-section.

Next, we will introduce definition of model of information retrieval of R.B. Yates and B.R. Neto [10]:

Definition 3.3:

A model of information retrieval is set of four $[D, Q, F, R(q_i, d_j)]$

in which:

+ D is a collection of document;

+ Q is a collection of user query;

ISBN: 978-1-941968-42-0 ©2017 SDIWC

+ F is a simulation framework of document, query and relation presentation;

+ $R(q_i, d_j)$ is a sorting function to link a real number with a query $q_i \in Q$ and a document representation $d_j \in D$. The sorting function determines order between documents and query q_i.

3. FUZZY INFORMATION SEARCH

3.1 Definition

Fuzzy search is searching information with incomplete input together with user's expectation or getting relative result with desired data. The reason we need studying and developing the fuzzy search:

+ User don't remember exactly searching term

+ Digital documents contain spelling errors when editing

3.2 Searching with Near Operator

In general, we search a character string by inputting exactly that string. For example, when we want to search information of Bill Gates, our character string query is "William Henry Gates". However, search engine will not give out pages which have the same information on Bill Gates, only include string "William H. Gates" or "William Gates". In order to solve this matter, we can use OR operator as follows: "William Henry Gates" OR "William H. Gates" OR "William Gates".

To replace using many OR operators as above and improve searching efficiency of desired Web, some search engines have developed NEAR operator. Herein, NEAR operator means searching pages contain nearby words. For example, "William NEAR Gates". How exactly "near" will depend on each specific search engine in the digital library.

For example: query "Digital Near/3 Library" in the IEEE digital library, search engine queries Web pages containing Digital and Library words not exceed 3 words. [7, 9]

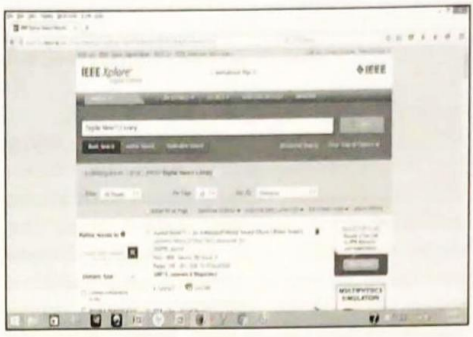

Fig. 1 – Operator NEAR in the IEEE digital library

3.3 Searching by Root Words

Progress of searching by root words is root-finding allows to spend less time on comparing query terms to index terms, for example both "computation" and "computer" are accepted to equal with "compute". Root-finding progress which eliminates one or more suffix(es) from word to reduce into root words, converts to neutral term without tense and plural state. The easiest way to evaluate exactly root-finding progress is considering some sample texts. In order to create index terms, all punctuations should be eliminated and as above discussion, all letters are grouped into word form.

At a glance, result of root-finding progress is not much value - the algorithm is not value. However, the first note, the last expression of root words is not important if it's unique to class of root terms and the second note, the alternation is repeated. Of course, it's not necessary to reverse the alternation because we only use root-finding progress to create index terms and texts are also stored exactly; the root word doesn't need to be any meaningful English word. In fact, "computer", "computing" and "computation" can be changed into "ppzxg" by root-finding progress if there is not any other word.

Root-finding progress is not necessary to fit with all elements of database. If there is a book in the list of library and putting the exact query "Arms AND Digital AND Libraries", root-finding progress will be mapped into one phrase with "Arms AND Digital AND

ISBN: 978-1-941968-42-0 ©2017 SDIWC

Library" and it may result in polysemy. In author field of directory database, it's suitable to disable root-finding progress; limit the root-finding progress with certain section of document which is optional, is available with database designer.

Actual installation of root-finding progress requires deep knowledge on language. For example, for English, in some cases: suffixes such as: -s, -ed, -ing, -ly and etc. are easy. In the other hand, exception of simple rule was recognized and supplemented by deep rules to prevent them being searched by root word; after that, they recognized other exceptions one by one etc. Root-finding progress contains more than 500 rules and exceptions, encoding as a limited state.

The main reason of root-finding and word-grouping progress is to simplify query formation. However, it causes significantly size of IF (Inverted File) becoming an advantage. There are 2 reasons: firstly, it has less stored IL (Inverted List) and they becoming denser, and storing price bases on one pointer is cheaper; secondly, average frequency of term inside the document tends to increase, it means that there is less pointer.

For example, after processing word-grouping and root-finding for TREC database, the number of pointers decrease about 16%, distinct terms decrease about 40% and total space which is used by Golomb encoding, decreases about 30%. However, the saving by using index of root-finding which is compensated up to a certain level equaling value of storing vocabulary by root-finding progress. A vocabulary which is not processed by word-grouping or root-finding, can be shared among compact element of document and sub-index. It can not be done by a root-finding progress. In TREC, vocabulary requires about 5 MB, so it's suitable for processing root-finding with amount of saving about 35 MB.

In a SF (Signatured File), effect of processing root-finding and word-grouping decrease the number of distinctive terms appearing in each record, so it's necessary to reduce width of signature for a given matching, hence accumulating the saving.

3.4 Searching Synonym Words

Any natural language has synonym words. When searching information on certain matter, we input typical keywords and get results from search engine. However, not only web pages contain query keywords including desired content, but also other web pages including these contents. In general, these web pages contain synonym words with query keywords. Hence, they studied and developed searching with synonym words of search engine.

Thus, searching with synonym words of search engine building set of synonym words dictionary according to supported language. When searching, search engine will find all web pages containing synonym words with query keyword. For example, when searching information on web crawler, search engine will find all web pages containing information on web robot, spider etc.

The advantage of searching synonym words is that users don't need to remember all synonym words. Especial in new technology fields, there are many terms used for a same matter that users can not know about it.

4. CONCLUSION

Nowadays, DL becomes important in national and international aspects because of information explosion follow by exponential function on Web. Web interface develop from browsing to searching, everybody do searching on Web everyday in the world. Thus, they have been concentrating on studying and developing technologies of searching in big database. There are many databases distributed all over the world where each of small group maintains document database by oneself. In DL, not only searching text information, but also searching multimedia information, search engines have been studied and improved, in which including method of retrieving fuzzy information.

APPENDIX: [8]

ISBN: 978-1-941968-42-0 ©2017 SDIWC

Algorithm of Searching Fuzzy Information

```
Search (T,n,p,m,k)
{
        /* Processing */
For each c ∈ Σ
{
T[c] = (c != p[m]) (c != p[m+1])
... (c != p[1]);
T[c] = 0s_{k+1}(t[c], 0) 0s_{k+1}(t[c],1)
...
        0s_{k+1}(t[c], m-k-1)
S[c] = ( c ∈ p[1..k+1]);
}
D_in = (0 1^{k+1})^{m-k} ;
M1 = (0^{k+1} 1)^{m-k} ;
M2 = (0^{k+1} 1)^{m-k-1} 0 1 1^{k+1} ;
M3 = 0 (k+1 0 1^{k+1} ;
G = 1 << k ;
/* Searching */
D = D_in ;
i = 0 ;
 while ( ++i <= n)
        if  (S[T[i]]) do
        {x  =  (D  >>  (k+2))  |
S[T[i])
        D = (( D << 1) | M1 ) &
((   D   <<    (k+3))    |    M2    )
        & (( x + M1 ) Λ x ) >> 1
) & D_in
        if  ( D & G == 0)
        {
                D = D | M3
        }
}
while    ( D != D_in && ++i <=n )
```

REFERENCES

[1] Arms W.Y. (2003), Digital Libraries, MIT Press, Cambridge.

[2] Chowdhury G.G. (1999), Introduction to Modern Information Retrieval, Library Assocciation Publishing, London.

[3] Lesk M. (2005), Understanding Digital Libraries, 2nd Edition, Morgan Kaufmann, San Francisco.

[4] Large A., Tedd L.A., Hartley R.J. (2001), Information Seeking in the Online Age, K.G. Saur Verlag, Munchen.

[5] Korfhage R.A. (1997), Information Storage and Retrieval, John Wiley, New York.

[6] Kowalski G. (1997), Information Retrieval Systems, Kluwer Academic Publishers, Boston.

[7] Schatz B.R., "Information Retrieval in Digital Libraries", Science 275, 1997, pp. 327-334.

[8] Wiederhold G. (2001), Database Design, 2nd Edition, McGraw-Hill, New York.

[9] Nguyễn Như Phong (2005), Lý thuyết mờ và ứng dụng, Nxb Khoa học và Kỹ thuật, TP. Hồ Chí Minh.

[10] Đỗ Quang Vinh (2009), Thư viện số: chỉ mục và tìm kiếm, Nxb Đại học Quốc gia Hà Nội.

ISBN: 978-1-941968-42-0 ©2017 SDIWC