Semantic keyword search based on information entropy

Ying Lou¹, Feng Zhong¹, JinXiang Zhang¹, Qian Li², *

¹School of Science and Technology, Zhejiang International Studies University, Hangzhou 310012, China
²Faculty of Economics and Law, Mogilev Sate A. Kuleshov University, Mogilev, 212022, Belarus

*Corresponding author: Jessica868816@outlook.com

Abstract. Keyword search is still the most effective means for users to obtain information. In the era of massive data, more and more structured and semi-structured data can be directly accessed by users. Different from the web data and text data, structured data and unstructured data contain a lot of semantic information, which can improve the accuracy of search results. Based on the theory of information entropy, we calculates the information entropy of semantic information to determine the semantic relevance of keywords, and then uses the degree of relevance between keywords and semantics, finally constructs a semantical model based on information entropy. Final experiments prove the effectiveness of our model.

1. Introduction
Keyword search is still one of the most friendly ways for people to get information. Customers still get web information from Google, Baidu and other webs. More and more customers will obtain information in specific fields through keyword search on specific data platforms (such as Alibaba, Weibo, Facebook). Different from text information and web information, specific information contains semantic information to information need. Figure 1 is an example of describing book information. Figure 1 (a) is an XML data fragment, and figure 1 (b) is book information in a relational database. Different from text information and web information, information fragments contain the semantic information of keywords. For example, "XML" is the name of a book chapter, "keyword" is the information of a reference data, "Tom" is the name of an author. If you use query language (SQL or XQuery), you can utilize the semantics, that should affect the friendliness of the query. Traditional keyword search methods such as BM25 model fail to consider these semantics in data.

2. Related works
In general, the information fragment in Figure 1 can always be expressed as structured information. For example, the information fragment (a) in Figure 1 can be represented as the structure information in Fig. 2, and the XML data is usually deemed as a tree. The information fragments in the relational database should be represented as graphic structure. Taking the tree structure as an example, we assume that users input keywords to query {xml, keyword, search} to query the books about XML keyword search, and to find the lowest common ancestor LCA [1] [2] which contains keywords is a
common method in such research. Among the existing research methods, LCA which contains key nodes in the tree structure is the basic method to obtain the query results of XML keywords. The calculation process is to extract the Dewey [3] coding list containing key nodes. The coding order of each keyword corresponds to the sequence of keywords appearing in the document. The calculation of Dewey coding can be used to get that the book of node 2, 11 and 16 is LCA with all keywords. In this paper, T(n) represents the subtree with n as the root node. In the above example, according to LCA concept, T(2), T(11) and T(16) will be returned as query results, and the output results are in the order of T(2), T(11) and T(16).

The researchers have improved the notion of LCA, and put forward the concept of SLCA (smallest LCA) [4]. According to SLCA, T(16) will be returned as the query result, and it can be seen that T(16) is not enough as the query result; The concept of interconnection is defined in literature [5] to eliminate the information fragments containing two or more identical label nodes in the query results. According to the concept of interconnection, T(15) will be excluded from the query results.

The paper proposes the ambiguity of XML keyword query and designs a keyword correlation ranking algorithm. The algorithm is divided into two steps: firstly, the relevance degree of elements and keywords is determined according to statistical information, and then the query target type of users is deduced. Based on the determination of the target type, the paper sorts the query results by using the proposed XML TF (term frequency) * IDF (inverse document frequency) weighting technology. In text retrieval, the maturity of TF * IDF model has been recognized, but in semantic data, the content of text information cannot play a greater influence in the sorting.

![Figure 1. Semantic information fragment](image1)

![Figure 2. The tree structure of XML fragment](image2)
The content information and structure information are considered separately in the literature [7]. The correlation degree of content information is calculated by TF * IDF model, and the compactness of the query results is used to calculate the correlation degree of the structure information. The two parts together constitute the final correlation of the query results. The relevance of query results of XML keywords is calculated from the content of XML, the level of XML hierarchy and the correlation degree between keywords.

From the above research, most of the researches focus on structure information and text information, and will use simple rules to judge the rationality of semantics. In some cases, it often affects the quality of query results.

3. Information entropy

3.1. Information entropy of semantics

In information theory, information entropy measures the uncertainty of random variables. In keyword query, the semantics of keywords are also uncertain. For example, in Figure 2, the semantics of keywords appearing under different tags are not same. The keyword "XML" may be the name of a book or the name of a data reference. For the keyword "Bob", it may be the author of a book or the name of a book. T (2), T (11) and T (15) belong to the same kind of information. In this kind of information, keywords may contain three aspects of semantics, which may represent the name of a book, the name of a chapter and the name of a reference.

**Definition 1: Semantics of Keywords**
The tag sequence of the keyword is the semantics of keyword, denoted as S, S= tag1/tag2/.../tagn.

According to definition 1, it can be found that the semantics of "search" in Figure 2 is bookstore/book/title, and the semantics of keyword "IR" is bookstore/book/chapter/title. A single document or a group of documents contains different semantics. The tag sequences are also different. Same semantics may contain different keywords. According to the information entropy calculation method, the distribution of keywords determines the value of a semantic information entropy.

**Definition 2: The Information of Entropy Semantics**
For the semantic s of a keyword, the information entropy is \( H(S) = -\sum_{x\in S} p(x) \log_2 p(x) \), where \( x \) is any keyword contained in semantics \( s \), \( p(x) \) is the probability of \( X \) appearing in semantic \( s \), \( p(x) = \frac{\text{The number of semantics } s \text{ that contained } x}{\text{The number of semantics } s} \).

According to definition 2, Table 1 lists the information entropy of partial semantics in Figure 2.

| Semantics          | Keyword | P(X) | H(S) |
|--------------------|---------|------|------|
| bookstore/book/title | XML     | 1/3  | 1.44 |
| bookstore/book/title | Bob     | 1/3  | 1.44 |
| bookstore/book/title | keyword | 2/3  | 1.44 |
| bookstore/book/title | Search  | 1    | 1.44 |
| bookstore/book/author | Bob    | 2/3  | 0.9  |
| bookstore/book/author | Tom    | 1/3  | 0.9  |

According to the theory of information entropy, the semantics with high information entropy contains more information. Compared with the examples in Table 1, the information of book / book / title is richer than that of book / book / author.

3.2. Information entropy of query results

The information need of keywords search has certain randomness. It is usually necessary to judge the relevance between keywords and search results. TF * IDF model has been successful in the field of text and web. SQL language and XQuery language can more clearly express the user's semantic needs. The
query results of SQL language and XQuery are used to eliminate the ambiguity of candidate results. It can increase the certainty of query results.

According to the content of the previous section, we can define the information entropy of the query results. The smaller the information entropy is, the greater the uncertainty will be. Regardless of using other means to guess semantics, the smaller the information entropy of the query results is, the more likely they are to be the query results that users want.

**Definition 3: Information entropy of search results**

For the result t of keyword query q = \{K1, K2, ..., Kn\}, its information entropy is \( H(T) = - \sum_{k \in Q, s \in T} H(s) * S(k) \). Where k is any keyword contained in query Q, and H(s) is the information entropy of semantic s. S(k) indicates whether the semantic s contains the shutdown key K, if the keyword K is included, the value of S(k) is 1. If the keyword K is not included, the value of S(k) is 0.

4. Keyword query based on information entropy

4.1. Ranking function

Utilizing the theory of information entropy, we measure the uncertainty of information fragments in XML documents. The results with smaller information entropy are preferred. With the help of keyword search technology used in text and web documents, we use TF * IDF model to measure the relevance of text information, and finally use formula 1 to measure the relevance of keywords and query results.

\[
F(Q, t) = M1 \times \frac{H(t)}{\max[H(T)] - \min[H(T)]} + (1 - M1) \times \sum_{k \in Q} \frac{TF(t(k)) \times IDF(T(k))}{\max(t(k))}
\]

(1)

For keyword search Q, the relevance of query result t consists of two parts. M1 is used as a parameter to adjust the proportion of information entropy in calculating the relevance. The value is between 0.5 and 1, which is determined by the richness of text information of query object. TF(t(k)) represents the frequency of keyword K in the query result T, and IDF(t(k)) represents its reverse word frequency. In order to balance the correlation score, the values of the two parts are normalized.

4.2. Data preprocessing

We need to preprocess the query data before executing queries. Following works should be done: a) we calculate the semantic information entropy of the query documents. According to the definition (2), semantic information entropy has nothing to do with query keywords, and each query semantic information entropy can be obtained by traversing the data; b) build the index of keyword inverted lists, which is stored by B+. For each keyword, the form of tuples in the inverted list of keyword k, each term is in the form of a tuple (Sid, Evalue, TFvalue). Sid represents the path that contains k. Evalue is the value of information entropy that k locates in the Sid, which can be calculated by the definition 2. TFvalue is the term frequency in path Sid.

5. Experimental Study

Comprehensive experiments have been conducted to evaluate the quality of our approach in this section. The empirical study is conducted on two real XML datasets: DBLP and IMDB. DBLP is a computer science bibliography with the size of 127MB. IMDB is the dataset of movies.
Comprehensive experiments have been conducted to evaluate the quality of our approach in this section. The empirical study is conducted on two real XML datasets: DBLP and IMDB. DBLP is a computer science bibliography with the size of 127MB. IMDB is the dataset of movies.

The information entropy of semantics in dataset IMDB is shown in table 2. The values of “/Movie/Rating/Distribution” and /Movie/Directors/Director/Name are the most smallest in IMDB, which means that the keywords matching them are certainly. Those results should be return prior.

We compare the performance of our algorithm named IES with that of TopX approach and original TF*IDF. As can be seen from Figure 3, in the case of big data, our algorithm is better than the other two algorithms. If the same keyword appears in different semantic information, the algorithm ies will select the query result with less information entropy, that is, the query result with less ambiguity, while other methods will ignore this information.

### Table 2. Part information entropy of semantics in dataset IMDB

| Semantics                          | Information entropy |
|------------------------------------|---------------------|
| /Movie/Title                       | 0.836               |
| /Movie/Actors/Actors/Name          | 0.696               |
| /Movie/Rating/Distribution         | 0.453               |
| /Movie/Directors/Director/Name     | 0.513               |
| /Movie/Editors/Editors/Name        | 0.608               |
| /Movie/Composers/Composer/Name     | 0.565               |

### Figure 3. Precision of dataset (DBLP and IMDB)

6. Conclusion

Basing on the concept of information entropy, we propose a novel semantic ranking scheme of keyword search. The higher the information entropy that results have, the higher the ambiguity of the query results contain. For the first time, we apply the theory of information entropy to retrieval. The final experimental results show that our method has obvious advantages in discerning the ambiguity of information.

References

[1] D. Harel and R. E. Tarjan. Fast algorithms for finding nearest common ancestors. In SIAM J. Comput. 13(2), pages 338–355, 1984.

[2] B. Schieber and U. Vishkin. On finding lowest common ancestors: Simplification and parallelization. In SIAM J. Comput. 17(6), pages 1253–1262, 1988.
[3] Lu, J., Ling, T.W., Chan, C.Y., Chen, T.: From region encoding to extended dewey: On efficient processing of xml twig pattern matching. In: VLDB. (2005) 193-204.

[4] Xu, Y, Papakonstantinou, Y.: Efficient keyword search for smallest lcas in xml databases. In: SIGMOD. 2005 537-538.

[5] S.Cohen, J.Mamou, Y.Kanza and etc. XSEarch: A Semantic Search Engine for XML, In Proceedings of VLDB, 2003.

[6] Z. Bao, T. Ling, B. Chen and J. Lu. Effective XML Keyword Search with Relevance Oriented Ranking. in ICDE 2009, 517-528.

[7] G. Li, B. C. Ooi, J. Feng, J. Wang, and L. Zhou, Ease: Efficient and adaptive keyword search on unstructured, semi-structured and structured data, in SIGMOD, 2008, 903-914.

[8] Guoliang Li, Chen Li, Jianhua Feng, Lizhu Zhou: SAIL: Structure-aware indexing for effective and progressive top-k keyword search over XML documents. Inf. Sci. 179(21): 3745-3762 (2009).