A Document Similarity Measurement without Dictionaries

Chung-Chen Chen

Institute of Information Science, Academia Sinica,

Taipei, Taiwan, R.O.C.

E-mail : johnson@iis.sinica.edu.tw

Wen-Lian Hsu

Institute of Information Science, Academia Sinica,

Taipei, Taiwan, R.O.C.

E-mail : hsu@iis.sinica.edu.tw

Abstract   Document similarity measure is an important topic in information retrieval and document classification systems. But the document similarity measure of English cannot apply directly to Chinese. The percentage of English vocabulary covered by an English dictionary is very high, but it is much lower in the case of Chinese. In this paper we proposed a measure called *common keyword similarity (CKS)* that does not rely on dictionaries. The *CKS* measurement is based on the common substrings of two Chinese documents. Each substring is weighed by its distribution over document database. Experiment shows that weighting function has a great influence on the recall/precision evaluation, and the *CKS* measurement without using any dictionary is better than a system that uses dictionaries.

1. Introduction

The measurement of document similarity is an important topic in interactive information retrieval systems and automatic classification systems. One of the most popular measures is the *Cosine Coefficient (CC)* based on the *Vector Space Model (VSM)* [1][2][3]. However, the vector space model needs a dictionary to build vectors. This approach is not appropriate for Mandarin Chinese. There are many words that cannot be found in ordinary Chinese dictionaries, such as terminologies, people's names and place names. The problem is further aggravated by the fact that there are no delimiters between these words in Chinese. To overcome that problem, we proposed an alternative measure that does not require any dictionary.

In this paper, we proposed a document similarity measure based on the common substrings of two documents. Our approach looks for all maximal common substrings using a PAT-tree [4][5][6], which is a tree that keeps all substring-frequency information. The importance of each substring (or keyword) is then evaluated by its *discriminating effect*, which indicates how well the keyword fits the pre-defined classification. The summation of keyword importance after normalization is considered as the similarity between two documents.

2. Extracting common keywords from two documents by a PAT-tree

Our approach to measure document similarity is based on maximal common substrings. A data structure called a PAT-tree can be used to find such substrings. A PAT-tree is a binary "trie". The "trie" is a tree in which each internal node is split at the first byte that is different from its brothers. Figure 1 shows a compressed trie, which illustrates the encoding of Chinese sentence "張先生比李先生高"
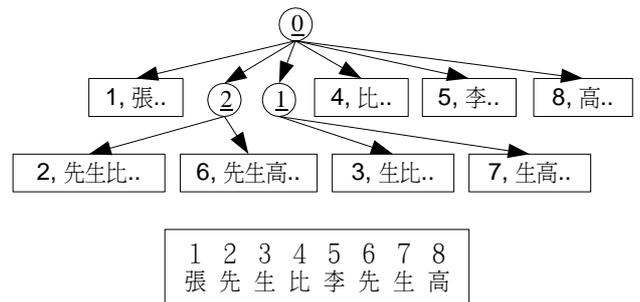

**Figure 1 : The compressed trie of Chinese sentence --
"張先生比李先生高"**

Because there are two substrings "先生" in the sentence of Figure 1, the "trie" split at the second child of the root. The number "2" in the node indicates that the two substrings have the first two characters in common.

A compressed binary trie has at most two children in each internal node. It keeps the indices of a text in linear space. A PAT-tree is a compressed binary trie that keeps the leaf node pointer and the split position in the same node, saving the space of leaf nodes. Figure 2 show a compressed binary trie and its PAT-tree after scanning the first six bits of the binary sequence "01100100010111".
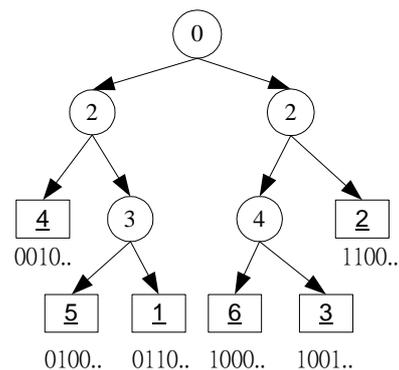

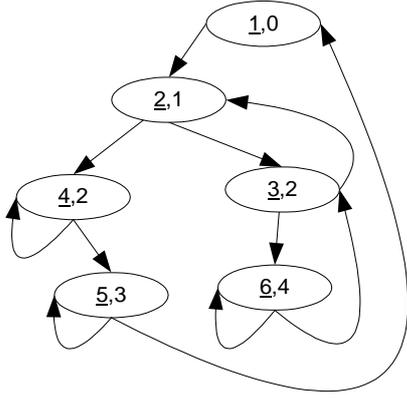**Figure 2a : Compressed Binary Trie**

Figure 2b : PAT-tree

The following algorithm illustrates how to get the maximal common substrings from two documents using a PAT-tree.

```
Algorithm get-common-keyword
Parameter doc1, doc2 : document
    PAT-tree1 = build-PAT-tree(doc1)
    pos = 1
    keyword-set = empty
    while pos <= length(doc2)
        tail2 = tail(doc2,pos)
        node1 = Search(tail2)
        tail1 = tail(doc1, node1.pos)
        keyword = longest-common-head(tail1, tail2)
        If keyword longer than 2 Chinese character
            Put keyword into keyword-set
return keyword-set
```

3. Measuring similarity between two documents

The measure of document similarity is normally based on the keyword weighting function. For an information retrieval system, some keywords are important while some are not. The situation also varies with different contexts. Stop words are good examples of keywords with no importance. A keyword may be important for some queries but not for others. Similarly, a keyword may be important in one classification but not in another.

In order to illustrate the idea of our keyword importance measurement, we shall introduce two classifications systems used in the news database of Central News Agency (CNA) at Taiwan. Each of them has four categories. Classification A is "Politics / Economics / Society / Education". Classification B is "Taiwan / Mainland of China / Overseas / Local". These two classification systems are some what orthogonal. In our measure, "Stock" is an important keyword in classification A because stock plays an important role in the economic market. But it is not important in classification B. On the other hand, the keyword "Russia" is important in classification B but not in A, because "Russia" is a foreign country for Taiwan.

In order to measure the importance of a keyword, two measures (*CE* and *KCE*) based on the distribution of documents over a classification are proposed. The ratio of these two measures reflects the importance of the keyword. In the following five definitions, the first four are proposed by us. The last definition *CC* is a popular document similarity measure that is used in contrast to our *CKS* measure.

Definition 1 : *Classification Entropy -- CE(CS)*
Let *D* be a set of documents. The classification system *CS* partition *D* into $D_1, D_2, ..., D_n$. Define $|D_i|$ as the number of documents in D_i. The *Classification entropy CE(CS)* is defined below.

$$CE(CS) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \left(\frac{|D_i|}{|D|} \right)$$

The entropy *CE(CS)* value is higher in a uniform distribution and lower in a biased distribution.

Definition 2 : *Keyword Classification Entropy--KCE(k, CS)*
Let D_k be the subset of *D* that contains the keyword *k*. The classification system *CS* partition D_k into $D_{1,k} D_{2,k} ... D_{n,k}$. The *keyword classification entropy KCE(k, CS)* is defined below.

$$KCE(k,CS) = -\sum_{i=1}^{n} \frac{|D_{i,k}|}{|D_k|} \log_2 \left(\frac{|D_{i,k}|}{|D_k|} \right)$$

KCE(k,CS) is the entropy of documents that contains keyword k over classification *CS*.

Definition 3 : *Keyword Discrimination Effect -- KDE(k, CS)*

$$KDE(k,CS) = \frac{CE(CS)}{KCE(k,CS)} - 1$$

The measure *KDE(k,CS)* reflects the importance of the keyword *k* over classification *CS*. A keyword is important for a classification if the documents containing the keyword are not evenly distributed within the classification system.

Definition 4 : *Common Keyword Similarity -- CKS(d_i, d_j)*
Let *len(d_i)* be the length of document d_i, w_k be the weight of keyword k. The *Common Keyword Similarity* of d_i, d_j is defined below.

$$CKS(d_i, d_j) = \frac{\sum_{k \in \text{maximal } common \text{ keyword of di, dj}} w_k^2}{len(d_i) * len(d_j)}$$

The measure *CKS* is based on the common substrings of two documents rather than a fixed dictionary. The weighting of each keyword w_k can be measured by any keyword weighting function such as $w_k=1$ or $w_k = KDE(k, A)$.

Definition 5 : *Cosine Coefficient -- CC(d_i, d_j) [1][2]*
Let w_{ik} be the weight of keyword k in document d_i, the *Cosine Coefficient* is defined below.

$$CC(d_i, d_j) = \frac{\sum_{k=1}^{L} w_k^2}{\sqrt{\sum_{k=1}^{L} w_{ik}^2 * \sum_{k=1}^{L} w_{jk}^2}}$$

The *CC* is very popular for document similarity measurement. It is used to make comparison to our *CKS* measure in the experiments.

4. Experimental results

The combination of recall and precision is a popular evaluation for information retrieval systems. It is defined as follows.

Definition 6 : Let D be a set of documents and d a target document. Let X be a human-defined subset of D that is similar to d, and let Y be the subset of D that is determined similar to d by program P. Then the *recall* and *precision* of P on document set D are defined as follows.

$$recall(P, d, D) = \frac{|X \cap Y|}{|X|}$$

$$precision(P, d, D) = \frac{|X \cap Y|}{|Y|}$$

Recall and precision are used to evaluate the quality of document similarity measures in the experiments. The subject corpus of our experiment contains 10056 news from *CNA*, March/1997.

We selected a total of 116 documents that contain keyword "陸委會" as the documents to be reviewed in our experiment. The first document with title "交通部將與陸委會商討九七台港航運" was selected as the target document. The experiments measured similarity between the target document and the collection of selected documents.

The recall/precision evaluation of the document similarity measure *CKS* is determined by several factors. The first factor is the keyword weighting function. In the keyword weighting function *KDE*, the classification plays an important role. Figure 3 shows the comparison between classification A : "Politics / Economics / Society / Education" and B : "Taiwan / Mainland of China / Overseas / Local". The recall/precision evaluation shows that classification A is much better than classification B in the experiment. This shows that classification B is not as important as classification A in judging the document similarity of news corpus.
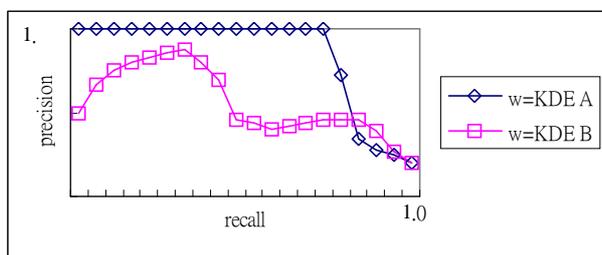

Figure 3. The effect of different classification on keyword weighting *KDE*

The second factor that influences the document similarity measure is the selection of keywords. Using dictionaries or not has a great influence on this measure. Figure 4 shows the comparison between *CKS* with and without dictionaries.
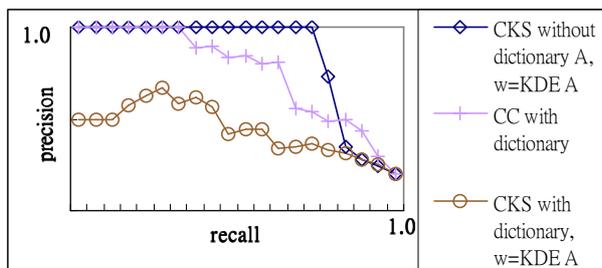

Figure 4. The effect of similarity measure with or without dictionary

The measure *CC* in definition 5 was used as a contrast measure in Figure 4. We used the dictionary collected from *Modern and Classical Chinese Corpora at Academia Sinica Text Databases*, which contains 78410 Chinese words [7]. The experiment shows that a system using common substrings is better than one using a dictionary. It is because many words in the news corpus cannot be found in an ordinary dictionary.

In these experiments, *KDE* exhibits very good performance except in the case of very high recall. It is mainly due to the fact that the target document has a low similarity to itself. A long common substring usually shows up only once in the corpus, which lower the weights of long keywords in weighting function *KDE*.

5.   Conclusion

A Chinese document similarity measure *CKS* based on common substrings and a keyword weighting function *KDE* are proposed in this paper. There are two observations. First of all, the document similarity measurement *CKS* without dictionaries is better than the *CC* measurement that uses a dictionary. Secondly, the classification system plays an important role in the keyword weighting function *KDE* for document similarity measurement.

Reference

[1] Salton, G., and Buckley, C. *"Term-Weighting Approaches in Automatic Text Retrieval,"* Information Processing & Management, 24, 513-23, 1988.

[2] Salton, G. *"Automatic Text Processing,"* Reading, Mass. : Addision Wesley, 1989.

[3] Rasmussen, E. *"Clustering Algorithm,"* Information Retrieval Data Structures & Algorithms, Prentice Hall, Edited by B. Frakes and Richardo Baeza-Yates, 419-442, 1992.

[4] Morrison, D., *"PATRICIA : Practical Algorithm to Retrieve Information Coded in Alphanumeric,"* JACM, 514-534, 1968.

[5] Gonnet, G. H., Baeza-Yates, R. A., Snider, T. *"New Indices for text : PAT trees and PAT arrays,"* Information Retrieval Data Structures & Algorithms, Prentice Hall, Edited by B. Frakes and Richardo Baeza-Yates, 66-82, 1992.

[6] Chien, L.F., *"PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval,"* The ACM SIGIR Conference, 50-58, 1997.

[7] Huang, C.R., Chen, K.J., *"Modern and Classical Chinese Corpora at Academia Sinica Text Databases for Natural Language Processing and Linguistic Computing,"* The Sixth CODATA Task Group Meeting on the Survey of Data Sources in Asian-Oceanic Countries, 9-12, 1994.