1C4-03

文書ベクトル列の複合汎化を用いた物語性に基づく文書検索

A Method for Detecting Narratives by Multi-Generalization of Document Vector Sequences

小木曽 聡*1 Akira OGISO　山本 章博*2 Akihiro YAMAMOTO

北海道大学工学研究科・知識メディアラボラトリ
Faculty of Engineering and MemeMedia Laboratory

In this paper we propose a new method for searching text files, with which we are aiming at finding similarity of narratives. We call text files documents in the paper. The method is applied after every document is transformed into a sequence of document vectors using the TF-IDF function. The main part of the method is the Length-Based Refinement (LBR, for short) mechanism for detecting similarity of pairs of document vector sequences. Before applying LBR, we amy generalize words by using a thesaurus. The generalization is called Generalization with the Concept Classification Dictionary (GCCD). In our system, unlike other existing systems, every query is given as a document, and then similarity based on is computed between the query document and those stored in a database. By interview to human beings we validated that our method reflects similarity of narratives.

1. Introduction

In searching text files we almost always depend on keywords, but often fail to find what we want. In this paper we propose a new method for search text files. We call text files *documents*. Our idea is making use of narratives described in documents. Narratives are very useful in managing information in our everyday life. When we read books, magazines, newspapers, or watch TV programs, we remember the contents as narratives. We expect that narratives help us search documents.

In this paper we regard narratives as semantics for documents. This means that we can neither store narratives in a database nor give a narrative as a query. In order to overcome the problem, we represent every document with an abstract data model, sequences of document vectors. We also introduce a similarity relation of the sequences with which we intend to detect similarity of narratives in documents. The similarity can be regarded as a type of "generalization" in the sense in inductive logic programming [Nienhuys-Chen 97]. We also use generalization of words by using a thesaurus. We validated by interview to human beings that our method reflects similarity of narratives.

2. Overview

In Fig. 1 we illustrate an overview of the system we are developing. The system is based on a relational database management system (RDMS). The RDMS stores corpora transformed into relations appropriate for the query processing. It also contains the Japanese Word Dictionary and the Concept Classification Dictionary developed and distributed by Japan Electronic Dictionary Research Institute (EDR) [EDR]. The Concept Classification Dictionary is a thesaurus, and the Japanese Word Dictionary gives a relation between words and concepts in the thesaurus.

In the system every document in a corpus is transformed into a sequence of document vectors and stored in a relation. Users give a query in the form of a document, which is also transformed into a sequence of document vectors. Then, the system searches documents in the corpora by applying a new generalization mechanism called Length-Based Refinement (LBR, for short) to pairs of document vector sequences. Before applying LBR, users can generalize words by using the thesaurus, the Concept Classification Dictionary. This generalization is called Generalization with the Concept Classification Dictionary (GCCD). All of these processes are executed with SQL queries.

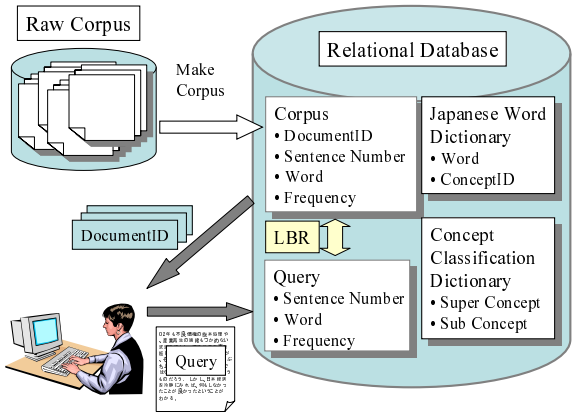

Figure 1: Overview of the Document Search System

3. Finding Similar Documents

We regard a *document* as a sequence of sentences, and a *sub-document* is a sub-sequence of the documents. We define a *narrative* as a sequence of events told in a document as shown in Fig 2. In a raw document one event is expressed in some sentences. However, it is difficult to find which sub-

連絡先: 山本章博, 北海道大学工学工学研究科・知識メディアラボラトリ, 〒 060-8628 札幌市北区北 13 西 8, Phone : 011-706-7253, Fax : 011-706-7808, Email : yamamoto@meme.hokudai.ac.jp

document explains exactly one event.

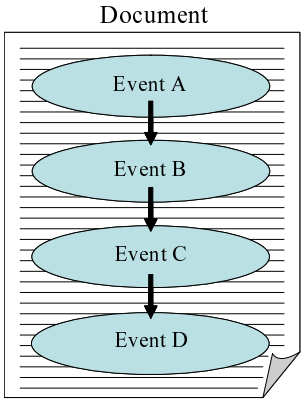

Figure 2: Narrative in a Document

We do not try detecting the exact sub-document for each event. Instead we generate sequences of sub-document from a document and compare it with a sequence of from another document.

3.1 Document Vector Model

In the document vector model, which words are appearing in a document is represented in the form of a vector. More formally, a document vector for a document d is $v(d) = (n_1, n_2, ..., n_k)$, where n_i is a value computed with the number of the occurrence of the word whose ID is i in the Japanese Word Dictionary. The element n_i in the vector is computed with the TF-IDF function $tf(t,d) \cdot isf(t)$, where $tf(t,d)$ indicates the frequency of the word t in the subdocument d, and

$$isf(t) = \log \frac{N(d)}{N(d,t)},$$

where $N(d)$ and $N(d,t)$ respectively denote the number of sentences of subdocument d and the number of sentences in d which contains t. Note that we regard a subdcoument as a "document" in the original definition of TF-IDF, and the whole document as a "corpus". When we need to distinguish such usage of TF-IDF to its original usage, we call the function *TF-ISF*.

In our research we treat documents written in Japanese, and therefore, we adopt Chasen, a tool for the morphological analysis [Matsumoto 00]. We also remove the stop words from them.

3.2 Length-Based Refinement (LBR)

The main part of our method is LBR. In LBR we make a window for each document d_i which indicates a subdocument w_i in d_i. Then, we compute the value $sim\big(v(w_1), v(w_2)\big)$, where sim is a function which defines similarity of two vectors. The pair of windows is slid from the heads of the each document to its ends simultaneously.

As the result we obtain a sequence of similarity value $S_i (i = 0, 1, ..., n)$ We conclude the similarity of d_1 and d_2 if

$$Pr(S_i > \tau) > 1 - \varepsilon,$$

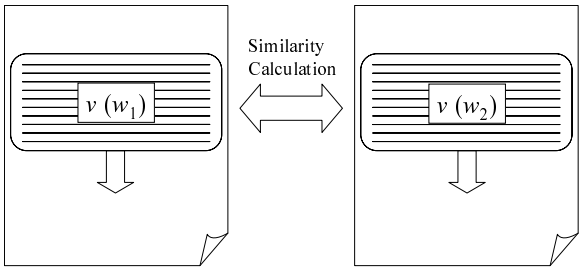

Figure 3: Length-Based Refinement

where τ and ε are parameters for the threshold and the error rate, respectively.

How long each window should be slid at each step is determined in the following manner: Let un assume that $N(d_1) \geq N(d_2)$. The window w_1 is slid for one sentence at every step. The window w_2 is slid for $\left\lfloor m_1 \times \frac{N(d_2)}{N(d_1)} \right\rfloor$ sentences, where $\lfloor x \rfloor = n$ represents the maximum integer satisfying that $x \geq n$.

The size of windows is given with a parameter $r (0 < r < 1)$ so that it holds $N(w_i) = \lfloor r \cdot N(d_i) \rfloor$ for $i = 1, 2$. The size must help us to recognize similarity of the narratives of the documents. In case that a window were too narrow, there would be almost no word which appears in common in the pair of windows, and therefore the similarity value will be very small even though the narratives of the two documents were similar. On the contrary, in case that the window size were too large, words appearing in common would affect the similarity value, without depending on where they occur in the documents.

We put a criterion for the value of r which is based on experiment with documents. Let a set $\{C_i\}_{i=1,2,...}$ of corpora and assume that every document in C_i tells a same narrative n_i, and that n_i is different from n_j if $i \neq j$. Then we require that, with r, any pair of d_i from C_i and d_j from S_j should be judged similar if and only if $i = j$. Moreover, we require that no pair of d and $s(d)$ should be judged similar, where $s(d)$ is obtained by changing the order of sentences in d randomly.

3.3 Generalization with the Concept Classification Dictionary (GCCD)

In our system users can generalize words with the Concept Classification Dictionary before applying LBR. The generalization is based on the a lattice which stores concepts in. If two words have a common super-concept, we replace them with the super-concept. The generalization is bounded by a parameter l called a generalization level. We can replace every concept c with its super-concept d which is in l steps beyond c. The generalization for the case $l = 0$ means that no generalization is applied to words. The generalization for the case $l = 1$ means that every word can be generalized to its super-concepts.

Because several super-concepts can be candidate as generalization of a word, we choose an appropriate super-

concept with the *fineness* F of a concept c defined as

$$F(c) = d(c) + \frac{1}{w(c)},$$

where the $d(c)$ is the *maximum* distance from the root to c and and $w(c)$ is the number of direct super-concept of c. Intuitively speaking, the larger $d(c)$ is, c is more detailed, and the smaller $d(c)$ is, c is more abstract. For two concepts c and d, when there is less super-concepts of c than d, we consider that c is more concrete than d. The super-concept is chosen according to the following rules:

R1 The word which is not registered in the Japanese Word Dictionary is not generalized.

R2 In case that exactly one common super-concept exists for several words, the words are generalized up to the common super-concept.

R3 In case that more than one common super-concepts exist, the words are generalized to the most detailed super-concept according to the fineness values of the concepts.

4. Experiment

At first we determined the value of each parameter used in LBR.

- For the similarity $sim(v, w)$ of two document vectors v and w. we adopt the cosine function

$$sim(v, w) = \frac{v \cdot w}{\|v\| \, \|w\|},$$

  where $\cdot$ indicates the inner product of two vectors, and $\| \ \|$ indicates the sizes of a vector, respectively.

- The value of the parameter r for window size was fixed with various books on narratives fairy tales like "Issunboushi", "Momotarou", and so on. The result is $r = 0.3$.

- Using another set of material documents, we concluded the threshold $\tau = 0.4$ and the error rate $\varepsilon = 0.35$.

The generalization level l for GCCD was fixed during the rather practical experiments reported below. In the experiment we found few differences of similarity of between similar documents and the others for the case $l = 0$. Both the similarity values between similar documents and these of other cases went up in the case $l = 2$. This means the difference of two cases became small. From these results, we concluded the appropriate $l = 1$.

We apply LBR and GCCD for document search from a corpus consisting of 2500 editorials in Mainichi Newspaper published during the period from April 1999 to December 2002. In the practice, if all the documents in the corpora were used as the document for retrieval, the computation time for finding similarity would become extremely long. Therefore, we use some keywords (for example, "near-miss", "accident", and so on) that decrease the candidate documents for retrieval, and choose the documents which contain the words and can be used as the target documents. After choosing the target documents, we choose one document as the query document, and we apply the LBR method between it and the rest of the documents.

For verification, we asked several persons to read documents which our system judged to be similar, and documents it did not so. At requesting to read them, we asked "Do you feel which documents have more similar narratives with the query document?". As the result, when we select "near-miss" and "accident" as keywords, seven of eight persons answered the same result as our retrieval system. When choosing "individual", "information", and "ledger" as keywords, six of seven persons answered the same way of the system. Although these are subjectivity evaluation, we can conclude that the documents considered by people to be similar have been retrieved by our system.

5. Conclusion

In this paper, we formalized narrative in a document as a sequence of events, and proposed a new method of searching documents whose narratives are similar. In our system, a query is given as a document and a relational database is used for supposing a lot of vectors and documents. We verified with people and concluded that coincides to human impression results obtained by our system.

References

[EDR] 日本電子化辞書研究所：EDR 電子化辞書 (現在は通信総合研究所が提供). http://www.jsa.co.jp/EDR/

[Haraguchi 02] M. Haraguchi, S. Nakano, and M. Yoshioka: Discovery of Maximal Analogies between Stories, *Proceedings of the 5th International Conference on Discovery Science (LNCS 2534)*, pp.324–331, 2002.

[Hearst 97] M. Hearst : TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, *Computational Linguistics*, 23 (1), pp. 33-64, March 1997.

[Kita et al. 02] 北 研二, 津田 和彦, 獅々堀 正幹: 情報検索アルゴリズム, 共立出版 (2002).

[Matsumoto 00] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, M. Asahara：*Morphological Analysis System ChaSen version 2.2.1 Manual* (2000). http://chasen.aist-nara.ac.jp/

[Nienhuys-Chen 97] Nienhuys-Chen and de Wolf : *Foundations of Inductive Logic Programming (LNAI Tutorial)*, Springer (1997).

[Tokunaga 99] 徳永 健伸: 情報検索と言語処理, 東京大学出版会 (1999).