Term Extraction Using A New Measure of Term Representativeness

Toru Hisamitsu Yoshiki Niwa Shingo Nishioka Hirofumi Sakurai
Osamu Imaichi Makoto Iwayama Akihiko Takano

Central Research Laboratory
Hitachi, Ltd.
Akanuma 2520, Hatoyama, Saitama, 350-0395, Japan
{hisamitu, yniwa, nis, hirofumi, imaichi, iwayama, takano}@harl.hitachi.co.jp

Abstract

This paper describes term extraction using a novel measure of term representativeness (i.e., informativeness or domain-specificity). The measure is defined by the normalized distance between the word distribution in the documents containing the term and the word distribution in the whole corpus. The measure works well in discarding uninformative frequent terms. We applied the measure to term extraction from abstracts of artificial intelligence papers. This paper introduces the measure and reports its effect in term extraction.

Introduction

In information retrieval (IR), the number of retrieved documents is often very large and this makes it difficult to grasp the contents of the documents. It is therefore helpful to know representative (informative or domain-specific) words in the documents. We have been developing an information retrieval system *DualNAVI*, which has a window displaying a viewgraph of representative words in the retrieved documents (Takano et al. 2000). Figure 1 shows an example viewgraph for the query "電子マネー(electric money)" using articles from the 1998 archive of the *Nihon Keizai Shimbun*, a Japanese financial newspaper. The displayed words are basically selected by *tf-idf* (Salton et al., 1973), and arranged in order of frequency (higher frequency words appear in the upper part of the viewgraph).

One problem is that uninformative words (such as "year," "month," and "one") often appear in the window although we use a stop-word list. In addition, the construction of a stop-word list has been *ad hoc* and (semi)-automatic construction is difficult. Another problem is that the difference of representativeness of words is not reflected enough. For instance, "read" and "cipher" are displayed equally. To resolve these problems, we developed a new measure for term[1] representativeness, which is effective in sorting words in the order of their representativeness (or non-representativeness), and it particularly works well in discarding "frequent non-representative" words (Hisamitsu et al., 1999a), which are to be listed in the stop-word list. The aim of this paper is to report the effect of the measure in term extraction.

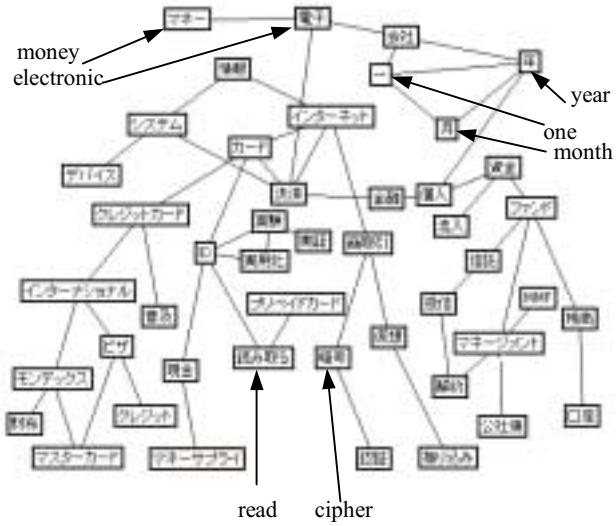

Figure 1
A sample view graph when the query is "電子マネー" (electronic money)

[1] We simply call a word or a word sequence a *term*.

1 Existing measures and their problems

1.1 Review of existing measures

Various methods for measuring the informativeness or domain specificity of a word have been proposed in the domains of IR and term extraction in NLP (see the survey paper by Kageura 1998). In characterizing a term, Kageura introduced the concepts of "unithood" and "termhood": unithood is "the degree of strength or stability of syntagmatic combinations or collocations," and termhood is "the degree to which a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts." Kageura's termhood is therefore what we call representativeness here.

Representativeness measures were first introduced in an IR domain for determining indexing words. The simplest measure is calculated from only word frequency within a document, For example, the weight I_{ij} of word w_i in document d_j is defined by

$$I_{ij} = \frac{f_{ij}}{\sum_k f_{kj}},$$

where f_{ij} is the frequency of word w_i in document d_j (Sparck-Jones et al. 1973, Noreault et al. 1977). More elaborate measures for termhood combine word frequency within a document and word occurrence over a whole corpus. For instance, *tf-idf*, the most commonly used measure, was originally defined as

$$I_{ij} = f_{ij} \times \log(\frac{N_{total}}{N_i}),$$

where N_i and N_{total} are, respectively, the number of documents containing word w_i and the total number of documents (Salton et al. 1973). There are a variety of definitions of *tf-idf*, but its basic feature is that a word appearing more frequently in fewer documents is assigned a higher value.

If documents are categorized beforehand, we can use a more sophisticated measure based on the χ^2 test of the hypothesis that an occurrence of the target word is independent of categories (Nagao et al. 1976).

Research on automatic term extraction in NLP domains has led to several measures for weighting terms mainly by considering the unithood of a word sequence. For instance, mutual information (Church et al. 1990) and the log-likelihood (Dunning 1993, Cohen 1995) methods for extracting word bigrams have been widely used. Other measures for calculating the unithood of *n*-grams have also been proposed (Frantzi et al. 1996, Nakagawa et al. 1998, Kita et al. 1994).

1.2 Problems of existing measures

Existing measures suffer from at least one of the following problems:

(1) Classical measures such as *if-idf* are so sensitive to term frequencies that they fail to avoid very frequent non-informative words.
(2) Methods using cross-category word distributions (such as the χ^2 method) can be applied only if documents in a corpus are categorized.
(3) Most measures in NLP domains cannot treat single word terms because they use the unithood strength of multiple words.
(4) The threshold value for being representative is defined in an *ad hoc* manner.

The scheme that we describe here constructs measures that are free of these problems.

2 New measure of term representativeness

This section briefly introduces the novel measure following (Hisamitsu, et al, 1999a).

2.1 Basic idea

We used the following working hypothesis, which is an interpretation of Firth's famous quote "You shall know a word by the company it keeps." (Firth, 1957):

If a term T is representative, the distribution of words in all documents containing T should have a bias from the "average" word distribution.

Following this hypothesis, we defined a measure based on a "normalized" distance between two word distributions. Let us first define some basic notations:

T: a term.
$D(T)$: the set of all documents containing T.
D_0: the set of all documents.
$P_{D(T)}$: word distribution in $D(T)$.
P_0 : word distribution in D_0.
$Dist\{P_{D(T)}, P_0\}$: the distance of two distributions $P_{D(T)}$ and P_0.

There are several measures of the distance between distributions, such as the log-likelihood ratio (LLR), Kullback-Leibler divergence (KLD), transition probability (TP), and one based on the vector-space (or cosine) method (VSM). The definitions of these four measures are given in the Appendix. We tried all four measures, but here will discuss only the case of LLR, which gave the highest performance.

We then defined $Rep(T)$, the representativeness of T, as

$$Rep(T) = Dist\{P_{D(T)}, P_0\}/B(D_0, \#D(T)),$$

where $\#D(T)$, the size of $D(T)$, denotes the number of words in $D(T)$, and the normalizing factor $B(D_0, \bullet)$, which we call a baseline function, estimates $Dist\{P_D, P_0\}$ where D is a randomly selected document set such

that $\#D = \#D(T)$. Actual calculation is slightly different, but is done in essentially the same way. See subsection 2.3.

2.2 Necessity of normalization

Figure 2 explains the necessity of the normalization. Displayed words correspond to coordinates ($\#D(T)$, $Dist\{P_{D(T)}, P_0\}$)s where T varies over 暗号 (cipher), 年 (year), 月 (month), 読み取る (read), 一 (one), する(do), and 経済 (economy). Figure 2 shows that, for example, $Dist\{P_{D(\text{“する(do)”})}, P_0\}$ is smaller than $Dist\{P_{D(\text{“経済(economy)”})}, P_0\}$, which reflects our linguistic intuition, but $Dist\{P_{D(\text{“暗号(cipher)”})}, P_0\}$ is smaller than $Dist\{P_{D(\text{“読み取る(read)”})}, P_0\}$ and even smaller than $Dist\{P_{D(\text{“する(do)”})}, P_0\}$, which contradicts our linguistic intuition. This is why values of $Dist\{P_{D(T)}, P_0\}$ are not directly used to compare the representativeness of terms. This phenomenon arises because $Dist\{P_{D(T)}, P_0\}$ generally increases as $\#D(T)$ increases. We therefore need to use normalization to offset this underlying tendency. The baseline function $B(D_0, \bullet)$ was designed so that it approximates the "baseline" curve in the lower part of Fig. 2, which plots coordinates ($\#D$, $Dist\{P_D, P_0\}$)s where D varies over randomly selected document sets, and is used to normalize the values of $Dist\{P_{D(T)}, P_0\}$.

From the definition of the distance, it is obvious that $B(D_0, 0) = B(D_0, \#D_0) = 0$. At the limit when $\#D_0 \rightarrow \infty$, $B(D_0, \bullet)$ becomes a monotonously increasing function.

2.3 Calculation of baseline functions

The curve could be approximated precisely through logarithmic linear approximation near (0, 0). To make an approximation, up to 300 documents are randomly sampled at a time. (Let this be denoted by D again). The number of sampled documents are increased from one to 300, repeating each number up to five times.) Each ($\#D$, $Dist\{P_D,P_0\}$) is converted to ($\log(\#D)$, $\log(Dist\{P_D, P_0\})$).

The curve formulated by the ($\log(\#D)$, $\log(Dist\{P_D, P_0\})$) values, which is very close to a straight line, is further divided into multiple parts and is approximated part-wise by a linear function. For instance, in the interval $I = \{x \mid 10{,}000 \leq x < 15{,}000\}$, $\log(Dist\{P_D,P_0\})$ could be approximated by $1.103 + 1.023 \times \log(\#D)$ with $R^2 = 0.996$.

2.4 Calculation of baseline functions

Therefore we define $Rep(T)$, the representativeness of T by normalizing $Dist\{P_D, P_0\}$ by $B(D_0, \#D(T))$ as follows:

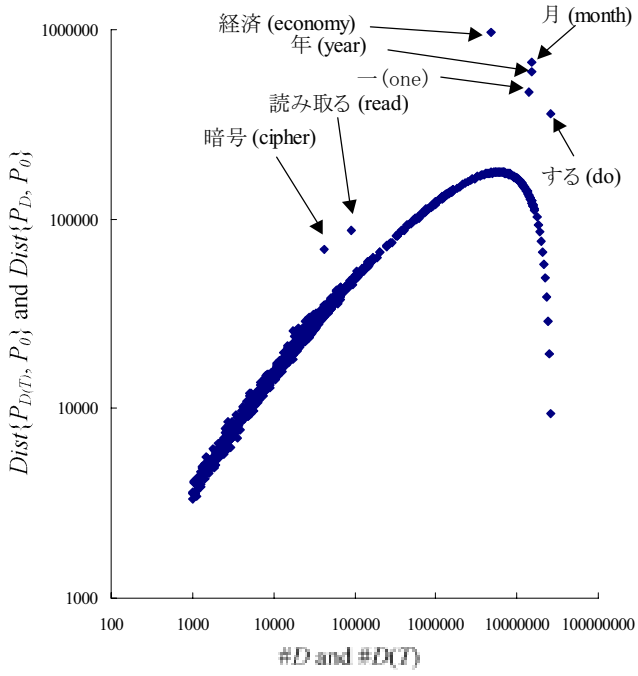

Figure 2
Baseline and sample word distribution

$$Rep(T) = 100 \times (\frac{\log(Dist(P_{D(T)}, P_0)}{\log(B(D_0, \#D(T)))} - 1).$$

For instance, when we used *Nihon Keizai Shimbun* 1996, for the randomly selected document set D, the average of $100 \times \log(Dist\{P_D, P_0\})/\log(B(D_0, \#D)) - 1)$, *Avr*, was 0.00423 and the standard deviation, σ, was about 0.465. Every observed value fell within *Avr*±4σ and 99% of observed values fell within Avr±3σ. This happened in all the corpora (7 corpora shown in Table 1) we tested. Therefore, we can define the threshold of being representative as, say, *Avr* + 4σ.

2.5 Treatment of very frequent terms

So far we have been unable to treat extremely frequent terms, such as する (do). We therefore used random sampling to calculate the *Rep*(*T*) of a very frequent term *T*. If the number of documents in *D*(*T*) is larger than a threshold value *N,* which was calculated from the average number of words contained in a document, *N* documents are randomly chosen from *D*(*T*) (we used *N* = 150). This subset is denoted $\underline{D}(T)$ and *Rep*(*T*) is defined by $100 \times \log(Dist\{P_{\underline{D}}, P_0\}) / \log(B(D_0, \#\underline{D})) - 1)$. This is effective because we can use a well-approximated part of the baseline curve; it also reduces the amount of calculation required.

2.6 Features of *Rep*(•)

By using *Rep*(*T*) defined above, we obtained *Rep*(する(do)) = 0.573, *Rep*(読み取る(read)) = 4.08, and *Rep*(暗号(cipher)) = 6.80, which reflect our linguistic intuition.

Rep(•) measure is free from the problems stated in subsection 2.1, and has the following advantages by virtue of its definition:

(1) Its definition is mathematically clear.
(2) It can compare high-frequency terms with low-frequency terms.
(3) The threshold value of being representative can be defined naturally.
(4) It can be applied to *n*-gram terms for any *n*.

Rep(•) measure has turned out to be effective in sorting words in the order of representativeness (or non-representativeness), and particularly works well in discarding "frequent non-representative" words (Hisamitsu et al. 1999a).

3 Term extraction method and its evaluation

3.1 Standpoint

The novel measure was originally developed to pick out representative terms from retrieved documents so that a user can grasp the topics in the documents. Therefore it primarily aims at eliminating frequent but un-informative terms, and finding "core" terms of medium-frequency. Rare terms are acceptable but they are not intended as the original targets. We mainly treated single-word and two-word terms for simplicity. The recall of our result was relatively low because of this simplification.

3.2 Text Corpora

We used parts of NACSIS Test Collection 1 (Kando et al. 1999): 1,870 abstracts of papers on artificial intelligence written in Japanese (NACSIS-AI), and the Japanese part of the NACSIS Test Collection 1

Table 1
Corpora and statistics on their content words

Corpus name	Description	# of total words	# of different words
NK96-50000	50,000 randomly selected articles from the whole corpus NK96 (206,803 articles of *Nihon Keizai Shimbun* 1996)	13,498,244	127,852
NK96-10000	100,000 randomly selected articles from NK96	26,934,068	172,914
NK96-158,000	15,8000 articles from *Nihon Keizai Shimbun* 1996 which do not include non-standard articles such as company personnel affairs	42,555,095	210,572
NK96-200000	200,000 randomly selected articles from NK96	53,816,407	233,668
NK98-158000	158,000 randomly selected articles from articles in *Nihon Keizai Shimbun* 1998	39,762,127	196,261
NACSIS-ALL	Japanese part of all abstracts (333,003 abstracts) in the NACSIS* test collection (see Section 3).	64,806,627	350,991
NACSIS-158000	158,000 randomly selected abstracts from NC-ALL	30,770,682	231,769

* The National Center for Science Information Systems

(NACSIS-J), which contains about 330,000 abstracts. NACSIS-J is a superset of NACSIS-AI. Terms were extracted only from NACSIS-AI and we conducted two main experiments. In experiment E1, we only used NACSIS-AI, and in experiment E2, we embedded NACSIS-AI into NACSIS-J, and used information of NACSIS-J when calculating measure values.

3.3 NLP issues

3.3.1 Morphological analyzer

We conducted word-based term extraction and used a Japanese morphological analyzer to segment sentences into words (Sakurai et al. 1999).

3.3.2 Grammatical filters

After the morphological analysis of a document, grammatical filters scanned every word and word bigram in the JMA output in order to eliminate obviously inappropriate items such as functional words, functional word bigrams, bigrams containing no nouns, etc.

3.4 Preliminary experiments on bigram selection

We conducted preliminary experiments on sorting word bigrams by using several measures (including MI, LLR, frequency, *tf-idf*, and the *Rep*(•) measure) and examined the top 100 bigrams from a qualitative viewpoint. In the experiments, the number of highly frequent non-representative bigrams, such as "**本-論文** (this paper)," appearing in the top 100 bigrams for each of the measures were shown in Table 2, where *tf-idf* was defined as follows so that it can be used to calculate a specificity value of a term against a whole corpus:

$$tf - idf = \sqrt{TF(T)} \times \log \frac{N_{total}}{N(T)},$$

where T is a term, $TF(T)$ is the term frequency of T, N_{total} is the number of total documents, and $N(T)$ is the number of documents that contain T.

In the case of MI, there were no frequent non-representative bigrams because all words were either too rare or too specific to be representative terms. This result indicates that the *Rep*(•) measure works well in discarding non-representative terms.

3.5 Combination of *LLR* and *Rep*(•)

In the experiment above, *Rep*(□) also picked out several non-AI-relevant terms because they were contained in an exceptional article about the economy. When we present extracted terms as an ordered list, AI-relevant terms should appear it is not preferable that irrelevant terms appear near the top of the list. On the other hand, it has been reported that frequency can be the most reliable measure in picking out "important" terms (Caraballo et al. 1999, Daille et al. 1994) because important core terms are actually used frequently, however, frequent non-representative terms cannot be discarded and this is a serious shortcoming. We therefore combined LLR, which retains frequency order well but works better than frequency, with *Rep*(•) to take advantage of the benefits of each: first terms are sorted by LLR and then those terms whose *Rep*(•)-values are under a threshold described in subection 2.2 are eliminated.

3.6 Treatment of longer terms

Each selected word bigram was examined as to whether it appeared independently or only as a part of a longer word sequence. In the latter case, we discarded the bigram and extracted every trigram which appeared independently and contained the discarded bigram, if the trigram's representativeness value was larger than the threshold.

3.7 Evaluation

We participated in a workshop on term extraction (Hisamitsu et al. 1999b), and our extracted terms were evaluated by the NACSIS Workshop TMREC Group (Kageura et al. 1999). Six other teams participated and they submitted 10 sets of extracted terms in total.

The extraction results were compared against two term sets: Manual-candidates (MC) and Index-candidates (IC). MC (IC) contains 8,834 (671) term candidates in total, which were manually selected by the TMREC Group. MC and IC were prepared as reference sets rather than as the "correct solution".

Table 3 shows the recall and precision rates of terms that fully matched some terms in the reference sets (bracketed numbers are the rank of each rate in the 12 (E1, E2, and 10 others) submitted term sets). The extraction result of E1 can be characterized as "lowest-recall and high-precision" against MC, and "lowest-recall and highest-precision" against IC. The extraction result of E2 can be characterized as "low-recall and high-precision" against MC, and "higher-recall and highest-precision" against IC.

Our method picked out core terms with high precision because it eliminated non-representative, highly frequent terms. When NACSIS-J was used, the recall rate increased while maintaining the high level of precision.

Table 2
Number of highly frequent non-representative bigrams appearing in the top 100 bigrams

Measure	Frequency	*tf-idf*	LLR	*Rep*(•)	MI
Number of highly frequent non-representative bigrams	23	14	13	4	0

Table 3
(a) Result of E1

Number of extracted terms	Comparison with MC		Comparison with IC	
	recall	precision	recall	Precision
943	4.90% (12)	45.92% (3)	21.01% (12)	14.49% (1)

(b) Result of E2

Number of extracted terms	Comparison with MC		Comparison with IC	
	recall	Precision	recall	Precision
5275	26.61% (8)	44.59% (4)	62.30% (3)	7.93% (2)

4 Future works

4.1 Unit of word co-occurrence

In the experiments stated above, we used an article or an abstract as the basic unit in our study of word co-occurrence. Although this worked well in term extraction from collections of newspaper articles or academic paper abstracts, we need to define other units in order to attempt term recognition from a different domain, for instance, a single book. In such a case, we may be able to use the "word window" or some logical units such as paragraphs. At the moment we have not carried out experiments using a unit other than an article or an abstract. Defining other co-occurrence units is an important issue to investigate.

4.2 Analytical modelling of baseline functions

The baseline functions play a crucial role in defining our representative measure. The approximation of a baseline curve can be done within a few minites by using a fast word-article association engine which we developed for information retrieval. Moreover, we have experimantally clarified the robustness of a baseline function against the difference of corpora, which means the baseline function is portable (Hisamitsu et al. 2000). However, we have not yet obtained an analytic model of the baseline functions. This model would be based on relatively easily obtainable statistics that reflect the word distribution of the whole corpus. Obtaining an analytic model of a baseline function is theoretically interesting and practically important.

Conclusion

This paper described a term extraction method that combines grammatical filters, the log-likelihood ratio, and a novel measure of term representativeness. Our experiments indicate that this method is effective in picking out core terms (such as encyclopedia entries) with high precision because it retains frequent terms but effectively eliminates non-representative terms. We plan to apply this measure to IR domain tasks such as the construction of a stop-word list for indexing, and weighting terms in document-similarity calculations.

Acknowledgements

We would like to express our gratitude to Prof. Jun-ichi Tsujii of the University Tokyo, for his insightful comments.

This project is supported in part by the Advanced Software Technology Project under the auspices of Information-technology Promotion Agency, Japan (IPA).

References

Caraballo, S. A. and Charniak, E. 1999 Determining the specificity of nouns from text. *Proc. of EMNLP'99*, pp. 63-70.

Church, K. W., and Hanks, P. 1990 Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics* 6(1), pp.22-29.

Cohen, J. D. 1995 Highlights: Language- and Domain-independent Automatic Indexing Terms for Abstracting, *J. of American Soc. for Information Science* 46(3), pp.162-174.

Daille, B. and Gaussier, E., and Lange, J. 1994 Towards automatic extraction of monolingual and bilingual terminology. *Proc. of COLING'94*, pp.515-521.

Dunning, T. 1993 Accurate Method for the Statistics of Surprise and Coincidence, *Computational Linguistics* 19(1), pp.61-74.

Firth, J. 1957 A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, Philological Society, Oxford.

Frantzi, K. T., Ananiadou, S., and Tsujii, J. 1996 Extracting Terminological Expressions, *IPSJ Technical Report of SIGNL*, NL112-12, pp.83-88.

Hisamitsu, T., Niwa, Y., and Tsujii, J. 1999a Measuring Representativeness of Terms, *Proc. of IRAL'99,* pp.83-90.

Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M., and Takano, A. 1999b Term Extraction Using A New Measure of Term Representativeness, *Proc. of the first NTCIR Workshop*, pp.475-481.

(www.rd.nacsis.ac.jp/~ntcadm/workshop/OnlinePro-ceedings/)

Hisamitsu, T., Niwa, Y., and Tsujii, J. 2000 A Method of Measuring Term Representativeness - Baseline Method Using Co-occurrence Distribution-, *Proc. of COLING 2000.* (to appear)

Kageura, K. and Umino, B. 1998 Methods of automatic term recognition: A review. *Terminology* 3(2), pp.259-289.

Kageura, K., Yoshioka, M., Tsuji, K., Yoshikane, F., Takeuchi, K., and Koyama, T. 1999 Evaluation of the Term Recognition Task, pp.417-434.

(www.rd.nacsis.ac.jp/~ntcadm/workshop/OnlinePro-ceedings/)

Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., and Hidaka, S. 1999 Overview of IR Tasks at the First NTCIR Workshop, *Proc. Of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition,* pp.11-44.

(www.rd.nacsis.ac.jp/~ntcadm/data/data-en.html, www.rd.nacsis.ac.jp/~ntcadm/workshop/OnlinePro-ceedings/)

Kita, Y., Kato, Y., Otomo, T., and Yano, Y. 1994 A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria, *Journal of Natural Language Processing* 1(1), pp.21-33.

Nagao, M., Mizutani, M., and Ikeda, H. 1976 An Automated Method of the Extraction of Important Words from Japanese Scientific Documents, *Trans. of IPSJ*, 17(2), pp.110-117.

Nakagawa, H. and Mori, T. 1998 Nested Collocation and Compound Noun For Term Extraction, *Proc.of Computerm'98*, pp.64-70.

Niwa, Y., Nishioka, S., Iwayama, M., and Takano, A., 1997 Topic graph generation for query navigation: Use of frequency classes for topic extraction. *Proc. of NLPRS'97*, pp.95-100.

Noreault, T., McGill, M., and Koll, M. B., 1997 A Performance Evaluation of Similarity Measure, Document Term Weighting Schemes and Representation in a Boolean Environment, *Information Retrieval Research* (Oddey, R. N. (ed.)), London: *Butterworths*, pp.57-76.

Takano, A., Nishioka, S., Niwa, Y., Iwayama, M., Hisamitsu, T., Sakurai, H., and Imaichi, O. 2000 *DualNAVI* – dual view interfaces bridges dual query types, *Proc. of RIAO 2000* (3), pp.19-20.

Sakurai, H. and Hisamitsu, T. 1999 A Data Structure for Fast Lookup of Grammatically Connectable Word Pairs in Japanese Morphological Analysis, *Proc of ICCPOL'99*, pp.467-471

Salton, G. and Yang, C. S. 1973 On the Specification of Term Values in Automatic Indexing. *Journal of Documentation* 29(4), pp.351-372.

Sparck-Jones, K. 1973 Index Term Weighting, *Information Storage and Retrieval* 9(11), pp.616-633.

Appendix:

Definition of the distance between two word distributions

Here we give the definition of log-likelihood ratio (LLR), Kullback-Leibler divergence (KLD), transition probability (TP), and a distance measure based on the vector-space or cosine method (VSM). Let $\{W_1,...,W_n\}$ be the set of all words and k_i and K_i be the frequency of a word w_i in $D(W)$ and D_0 respectively.

• LLR

The likelihood ratio (LR) for a hypothesis H is the ratio of the maximum value of the likelihood function over the subspace represented by the hypotheses to the maximum value of the likelihood function over the entire parameter space. Here H is the hypothesis that the word distribution in $D(W)$ is independent of the word distribution of D_0. LLR is the logarithm of the LR, which is represented as follows:

$$\sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(W)} - \sum_{k=1}^{n} k_i \log \frac{K_i}{\#D_0}.$$

• KLD

The Kullbck-Leibler divergence of the word distribution in $D(W)$ with respect to the backgroung distribution is definred as follows:

$$\sum_{i=1}^{n} \frac{k_i}{\#D(W)} \log \frac{\frac{k_i}{\#D(W)}}{\frac{K_i}{\#D_0}} = \sum_{i=1}^{n} \frac{k_i}{\#D(W)} \log \frac{\#D_0 k_i}{\#D(W) K_i}.$$

• TP

The transition probability of the word distribution in $D(W)$ from the background distribution is defined as follows:

$$\sum_{i=1}^{n} \sqrt{\frac{k_i}{\#D(W)}} \sqrt{\frac{K_i}{\#D_0}}.$$

The value is equal to 1 when the two distributions are identical to each other. Therfore we used

$$1 - \sum_{i=1}^{n} \sqrt{\frac{k_i}{\#D(W)}} \sqrt{\frac{K_i}{\#D_0}}$$

for the distance.

• VSM

The vector space model has been used in the domain of IR, and the distance of the two distributions is defined as

$$1 - \frac{\sum_{i=1}^{n} \frac{k_i}{\#D(W)} \bullet \frac{K_i}{\#D_0}}{\sqrt{\sum_{i=1}^{n} \left(\frac{k_i}{\#D(W)}\right)^2} \sqrt{\sum_{i=1}^{n} \left(\frac{K_i}{\#D_0}\right)^2}}.$$