# Term Extraction Using A New Measure of Term Representativeness

Toru Hisamitsu, Yoshiki Niwa, Shingo Nishioka, Hirofumi Sakurai,
Osamu Imaichi, Makoto Iwayama, and Akihiko Takano
Central Research Laboratory, Hitachi, Ltd.
{hisamitu, yniwa, nis, hirofumi, imaichi, iwayama, takano}@harl.hitachi.co.jp

**Abstract**
This paper describes a term extraction method that uses a novel measure to determine the representativeness (i.e., informativeness or domain-specificity) of a term. This measure is defined by normalizing the distance between the word distribution of the documents containing the query term and the word distribution of the whole corpus. The measure can compare the representativeness of two terms whose occurrence frequencies largely differ, and has a naturally defined threshold of determining whether a term is representative. We combined the measure with grammatical filters and extracted terms from abstracts of artificial intelligence papers with high precision.
**keywords: representativeness, term extraction, stop-word list**

## 1. Background

In information retrieval, the number of retrieved documents is often too large for a user to grasp the contents of the documents. It is therefore helpful to have an overview of the representative words in the documents for refining or expanding the query. To achieve this, one of the authors has been developing an information retrieval (IR) system called *DualNAVI*, which has a navigation window for displaying a viewgraph of representative words in the retrieved documents (Niwa 1997). Although the viewgraph has turned out to be quite helpful, it still has room for improvement.

Figure 1 shows an example viewgraph for the query "電子マネー(electric money)" (with a financial paper Nikkei Shinbun 1996 as the corpus). The words to be displayed are basically selected by *tf-idf* (Salton et al. 1973), and they are arranged in order of frequency (words with higher frequencies appear in the upper part of the viewgraph).

One problem with *DualNAVI* is that uninformative words often appear in the window. We used a stop-word list to suppress uninformative words (such as extremely common verbs or numerals), which greatly improved the appearance of the viewgraph (but , for instance, "上(on)" still appears in the graph though it is not very important as a keyword). However, construction of the stop-word list has been quite ad hoc and unsystematic. We defined the most frequently appearing words (for example, the top 2000 words) as the stop-words and added the words that had certain parts of speech (such as particles or auxiliary verbs) to them. Another problem is that the representativeness (informativeness or domain-specificity) of a word is not highlighted enough. For instance, "暗号化"(encryption) should be highlighted more than less representative words such as "読みとる"(read). To resolve these problems, we developed a new way of measuring the representativeness of a term (a word or a word sequence) that can be used to construct the stop-word list.

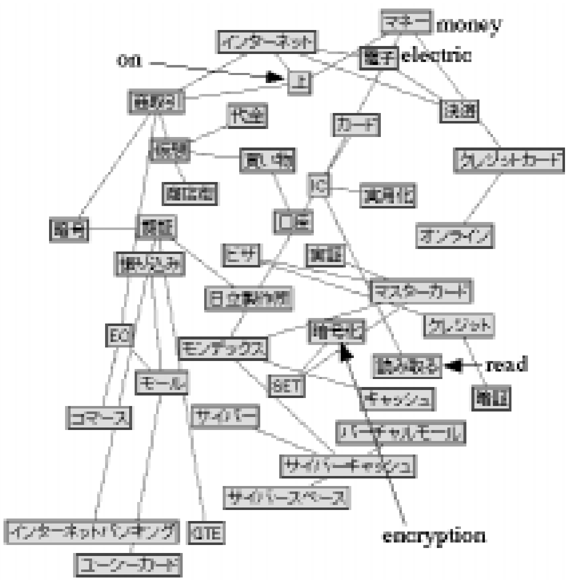


**Figure 1**
A view graph when the query is
"電子マネー (electric money)".

Section **2** reviews existing methods for measuring representativeness and points out their shortcomings. Section **3** introduces our measure for representativeness of terms. Section **4** describes a term extraction method which combines grammatical filters with the representative measure. Section **5** shows the qualitative and quantitative results of the term extraction method, and Section **6** concludes this paper.

## 2. Existing measures

### 2.1 Overview

Various methods have been proposed for measuring the "informativeness" or "domain-specificity" of a word in the domains of IR and term extraction. This section reviews the measures described in a survey paper on automatic term extraction (Kageura 1997). Kageura introduced the notions of *unithood* and *termhood*, which together characterize a term. Unithood is "the degree of strength or stability of syntagmatic combinations or collocations," and termhood is "the degree that a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts." Kageura's *termhood* is therefore very closely related to representativeness in this paper.

*Tf-idf*, the most commonly used measure for termhood, is defined by combining word frequency within a document and word occurrence over a whole corpus as follows:

$$f(w, d) \times \log(\frac{N_{total}}{N(w)}),$$

where $N(w)$ and $N_{total}$ stand for the number of documents containing word $w$ and the total number of documents, respectively. Although the definition of *tf-idf* has a number of variations, its basic feature is that a word appearing more frequently in fewer documents is assigned a larger value.

If the categories of the documents are known, we can apply a more sophisticated measure for termhood that is based on the $\chi^2$-test against the hypothesis that an occurrence of the target word is independent from the categories.

Research on automatic term extraction has been done in the domain of natural language processing (NLP), and several measures have been proposed for term weighting in term extraction. For example, mutual information (Church et al. 1990) and log likelihood (Cohen 1995) have been used to select word bigrams, and other measures have treated *n*-grams (Kita et al. 1994, Frantzi et al. 1994, Nakagawa et al. 1998).

### 2.2 Problems

Existing weighting measures have the following problems:

(1) Classical measures such as if-idf turned out to be ineffective. Without the ad-hoc stop-word list, the topic word graph of *DualNAVI* has quite a few non-informative words
(2) The methods for comparing cross-category word distributions (such as the $\chi^2$ method) can only be applied to a categorized document set.
(3) Most measures in NLP domains cannot be applied to single word terms.
(4) The threshold value for being important/ unimportant is often defined in an ad-hoc manner.

The next section gives a measure which is free from these problems.

## 3. A new representativeness measure

To begin with, let us restate the definition of representativeness from our standpoint. Since our purpose is to select terms for a navigation window, "representative" terms are informative terms that provide an overview of topics in the retrieved documents. Frequent but uninformative words and domain-specific but rare words are not our target.

### 3.1 Basic idea

Our basic idea can be summarized by the following famous quote (Firth 1957):

"*You shall know a word by the company it keeps*."

Let us give a straightforward mathematical interpretation of this phrase.

To begin with, let us introduce some basic notations:

$W$: a term, i.e., a word or a word sequence.

$D(W)$: the set of all documents containing $W$.

$D_0$: the set of all documents.

$P_{D(W)}$: word distribution in $D(W)$.

$P_0$ : word distribution in $D_0$.

We define $Rep(W)$ (the representativeness of $W$) that is based on $Dist\{P_{D(W)}, P_0\}$, the distance of two distributions $\{P_{D(W)}, P_0\}$. Normalization of the distance will be discussed in the next subsection.

There are several methods for measuring the distance between two distributions. They include log-likelihood ratio (LLR), Kullback-Leibler divergence, transition probability, and the vector-space or cosign method. We tried all four measures, but will discuss here only LLR, which gave the most stable results. $Dist\{P_{D(W)}, P_0\}$ is defined by using LLR as follows:

$$Dist(P_{D(W)}, P_0) = \sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(W)} - \sum_{i=1}^{n} k_i \log \frac{K_i}{\#D_0},$$

where $\{W_1,...,W_n\}$ is the set of all words, and $k_i$ and $K_i$ are the frequencies of a word $w_i$ in $D(W)$ and $D_0$, respectively.

The sample words displayed in Figure 2, which were randomly chosen from the *Nikkei-Shinbun* 1996, correspond to coordinates $(\#D(W), Dist\{P_{D(W)}, P_0\})$, where $W$ denotes a word, and $\#D(W)$ denotes the number of words contained in $D(W)$. The figure shows that

$Dist\{P_{D(\text{"する(do)"})}, P_0\}$ is smaller than $Dist\{P_{D(\text{"米国(USA)"})}, P_0\}$, which reflects our linguistic intuition. Similarly, $Dist\{P_{D(\text{"結び付ける(bind)"})}, P_0\}$ is smaller than $Dist\{P_{D(\text{"オウム(Aum)"})}, P_0\}$ as expected*.

However, as can be seen from the graph, $Dist\{P_{D(W)}, P_0\}$ increases as $\#D(W)$ increases, which means that direct comparison of $Dist\{P_{D(W_1)}, P_0\}$ and $Dist\{P_{D(W_2)}, P_0\}$ is inappropriate when $\#D(W_1)$ and $\#D(W_2)$ are considerably different. Consequently, $Dist\{P_{D(\text{"する(do)"})}, P_0\}$ is roughly equal to $Dist\{P_{D(\text{"オウム(Aum)"})}, P_0\}$, which is quite unnatural. We therefore need a kind of normalization.

### 3.3 Normalization of the distance

As stated above, direct comparison of $Dist\{P_{D(W_1)}, P_0\}$ and $Dist\{P_{D(W_2)}, P_0\}$ is problematic when two terms $W_1$ and $W_2$ have very different frequencies. Therefore, we studied the basic behavior of $Dist\{P_{\cdot}, P_0\}$, that is, the behavior of $Dist\{P_{\boldsymbol{D}}, P_0\}$ when $D$ is a randomly selected document set. The points marked by crosses in Figure 2 are ($\#D$, $Dist\{P_{\boldsymbol{D}}, P_0\}$)s where $D$ varies over sets of randomly selected document sets of various sizes. This figure indicates the existence of an underlying smooth curve, which we call a baseline curve. Its function is denoted as $B_{D0}(\cdot)$.

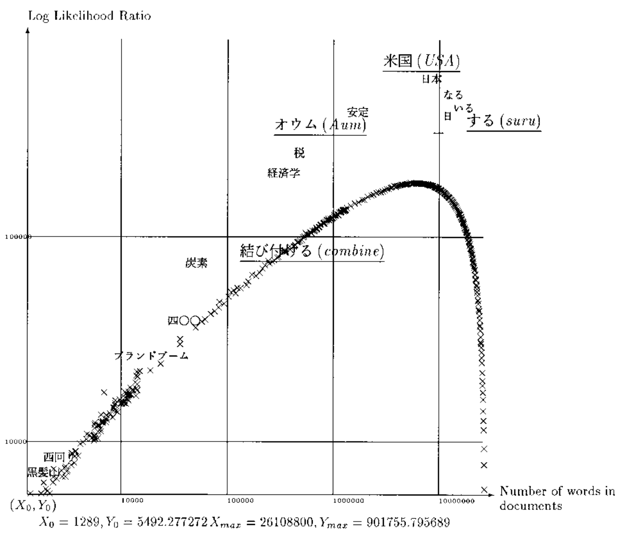


**Figure 2**

Baseline and sample word distribution

From the definition of the distance, It is trivial that $B_{D0}(0) = B_{D0}(\#D_0) = 0$. Note that (0, 0) is shared by every $B_{D0}(\cdot)$, while ($\#D_0$, 0) depends on an arbitrarily given $D_0$. In our experiments, the behavior of $B_{D0}(\cdot)$ is very stable and does not change very much around the origin when the size of $D_0$ is varied. $B_{D0}(\cdot)$ can be well approximated by a simple power function $B^*_{D0}(\cdot)$. In particular, in the interval $I = \{x \mid 1000 \leqq x < 20{,}000\}$, $B^*_{D0}(x)$ can be very closely approximated by a power function for various sizes of $D_0$ (from 2,000 documents to 300,000 documents).

Therefore, when $\#D(W)$ belongs to $I$, it is natural that $Rep(W)$ be defined as the normalization of $Dist\{P_{D(W)}, P_0\}$ by $B^*_{D0}(\cdot)$ as

$$Rep(W) = Dist\{P_{D(W)}, P_0\}/B^*_{D0}(\#D(\mathrm{W})).$$

In each of the experiments we conducted, the average of $Dist\{P_D, P_0\}/B^*_{D0}(\#D)$, Avr, was within 1.00(±0.01) and the standard deviation, σ, was about 0.05. Every observed value fell within Avr ± 4σ, which means that $B^*_{D0}(\#D)$ approximated $Dist\{P_D, P_0\}$ very well for randomly chosen documents. What is important here is that we can naturally define the threshold of a term being representative as, say, Avr + 4σ (≒ 1.2).

### 3.4 Treatment of very frequent terms and very rare terms

So far we have been unable to treat extremely frequent terms such as "する"(do). To resolve this problem, we used random sampling to calculate the $Rep(W)$ of a very frequent term $W$. If the number of documents in $D(W)$ is larger than a threshold value $N$, which is calculated from the average number of words that a document contain, $N$ documents are randomly chosen from $D(W)$. This subset is denoted by $\underline{D}(W)$ and $Rep(W)$ is defined by $Dist\{P_{\underline{D}(W)}, P_0\}/B^*D_0(\#\underline{D}(W))$. This method is advantageous not only because it uses the interval $I$, but also because it speeds up the calculation of $Rep(W)$.

The left-most value of the interval $I$ roughly corresponds to the number of words in three or four documents, and in the period $P = \{x \mid 0 \leqq x < 1000\}$, $B^*_{D0(\cdot)}$ has a tendency to overestimate $B_{D0(\cdot)}$. However, we simply used the $B^*_{D0(\cdot)}$ for a term $W$ with $\#D(W) <$ 1,000 because underestimating the weight of rare terms was harmless to our purpose.

### 3.5 Features of Rep(・)

$Rep(W)$ has the following favorable features:

(1) Its definition is mathematically simple and clear.
(2) It can compare high-frequency terms with low-frequency terms.

* “Aum” is the name of a rekigeous cult.

(3) The threshold value of being representative can be naturally defined.

(4) It can be applied to n-gram terms for any $n$.

The essential difference between the new measure and existing ones is that it treats the context of a target term in a sense as well as the distribution of the target term itself.

Results of the experiments on discrimination of informative/un-informative terms using $Rep(\cdot)$ will be reported elsewhere (Hisamitsu et al. 1999).

# 4. Description of term extraction method

This section describes the method for term extraction which combines our measure of representativeness and a set of grammatical filters.

## 4.1 Standpoint

The measure for representativeness was originally developed to pick out representative (informative or domain-specific) terms from a large number of retrieved documents so that a user can have an overview of the contents of the retrieved results. To solve the two problems stated in Section **1**, the measure mainly aims at eliminating very frequent but un-informative words, and finding medium-frequency core words, which represent a sizable but tractable number of documents.

Rare words (which appear in only a few documents) are passable but not the original targets of our term extraction. In *DualNAVI*, we are not planning to apply apply our measure to less frequent (frequency 3 or under) words in order that rare but characteristic words are not eliminated. Those words can be displayed in the navigation window, where words are classified into five classes according to frequency, and a part of words in each class are picked out to be displayed (Niwa 1997).

## 4.2 NLP techniques

・ **Morphological analyzer**

Because we conducted word-based term extraction, we used a Japanese morphological analysis (JMA) program called ANIMA to segment untagged corpora (1870 AI abstracts and NACSIS J-collection). The program has been described in detail (Sakurai et al. 1999).

・ **Grammatical filters**

We mainly treated one-word and two-word terms for simplicity. After the JMA program analyzed the articles in the AI abstracts, a set of grammatical filters scanned every word and every adjacent word pair in the JMA output to eliminate obviously inappropriate ones. For instance, the filter eliminated the following inappropriate words or word pairs:

(i) functional words (particles),

(ii) two-word sequences that contained no nouns,

(iii) two-word sequences whose first word was a functional word other than a nominal prefix, and

(iv) two-word sequences whose second word was a functional word other than a nominal suffix.

## 4.3 Selection by statistical measures

We prepared term-article matrices, which recorded which article contained which term how many times for all one-word and two-word term candidates. The matrices were used to calculate the representativeness of each term. The terms whose representativeness value were larger than 1.2 were selected.

## 4.4 Treatment of multi-word terms

Every surviving two-word term candidate was examined as to whether it actually independently appeared in an article or only as a part of a longer word sequence. In the latter case, we discarded the two-word term and extracted three-word term candidates which independently appeared and contained the two-word term candidate, and calculated their representativeness values. The criteria stated in 4.3 was applied to pick out three-word term candidates.

# 5. Results

## 5.1 Quantitative evaluation

We omit detailed discussion of the quantitative evaluation of the method described in Section 4 because it is already described in the "Candidate-evaluation" and "N-common evaluation" provided by NACSIS Workshop TMREC Group. We only briefly mention the results.

We could choose both the AI abstracts and the whole J-collection as the whole corpus $D_0$. It has turned out that using a larger corpus (J-collection) resulted in better performance in terms of recall (see the results of the method "b" in Fig. 3 of the TMREC evaluation), and using smaller one resulted in lower recall and higher precision (see the results of the method "a" in Fig. 3 of the TMREC evaluation). The effect of using a tagged corpus was slight (see the results of the method "f" in Fig. 3 of the TMREC evaluation).

What important is that the method seems to perform well in picking out core terms with high precision, while it eliminates highly frequent un-informative terms and has a tendency to neglect lower frequent terms.

## 5.2 Some qualitative results

To investigate the nature of our representative measure, we compared the top-100 two-word terms[**] of several statistical measures. We only compared two-word terms because the major portion of extracted terms were two-word terms, and we wanted to observe the effect of log likelihood ratio (LLR) (Dunning 1994) and mutual information (MI) (Church et al. 1990), which are frequently used measures for word bigrams. We also used *tf-idf* and frequency for comparison. Note that *Rep*(･), *tf-idf*, and frequency can be applied to word *n*-gram terms for any *n*.

Table 1, 2, 3, 4, 5, and 6 show the top-100 two-word terms extracted by using frequency, *tf-idf*, MI, LLR, *Rep*(･), and a combination of LLR and *Rep*(･) (first sort by LLR, and then eliminate un-informative terms by *Rep*(･)) respectively. Here *tf-idf* is defined as follows so that it can be used to calculate a specificity value of a term against a whole corpus:

$$tf\text{–}idf(W) = \sqrt{T(w)} \times \log(\frac{N_{total}}{N(w)})$$

where $T(W)$ is the total frequency of the term $W$ in the whole corpus.

As expected, mutual information worked very poorly because it overestimated low frequency terms. Frequency and *tf-idf* worked quite well, partly because it is natural to expect that important words would be used relatively often. Discarding frequently occurring unimportant words is therefore important.

In our experiments, 23, 14, 13, and 4 frequently occurring unimportant words[***] appeared in the top 100 words, when frequency, *tf-idf*, LLR, and *Rep*(･) were used for the extraction respectively. In the case of mutual information, there were no frequently occurring unimportant words, instead all words were too rare or too specific to be representative terms.

Table 5 contains several economical terms because the AI abstracts contained an exceptional abstract and *Rep*(･) sensitively picked out terms from it. However, they may seem to be irrelevant in the AI domain. To make the order more intuitively natural, we combined *Rep*(･) with LLR: first sorted terms by LLR and then eliminated inappropriate terms with *Rep*(･). Table 6 shows the result, which seems to be better than both LLR and *Rep*(･).

In general, *Rep*(･) is able to extract representative terms effectively when the corpus consists of similar-sized and similar-styled documents. It is very effective for discarding highly frequent unimportant terms. Thus *Rep*(･) is suitable for constructing stop-word lists.

Table 1

Top-100 terms when frequency is used.

| | | | |
|---|---|---|---|
| 本稿 594 | 動的 73 | 自然言語 55 | 解決過程 39 |
| 学習者 496 | 対象モデル 72 | 学習アルゴリズム 55 | 機械翻訳 37 |
| 問題解決 445 | 相互作用 72 | ベース推論 55 | 支援環境 36 |
| 本論文 420 | ＣＡＩシステム 71 | 定性推論 54 | 思考過程 36 |
| 本研究 390 | 音声認識 70 | 故障診断 54 | 最適解 36 |
| 知的 243 | 論理プログラム 69 | 因果関係 54 | 一階 36 |
| 知識ベース 229 | 類似度 69 | 強化学習 50 | 知識処理 35 |
| 支援システム 213 | 定式化 68 | 構造化 49 | 機能モデル 35 |
| 有効性 166 | 自動的に 68 | 設計過程 47 | 目的 34 |
| 本システム 142 | 推論システム 63 | 教材知識 47 | 対象世界 34 |
| 知識表現 133 | 時間 63 | 自動生成 46 | 多項式時間 34 |
| 知識獲得 127 | 決定木 62 | 学習環境 46 | 述語論理 34 |
| 再利用 100 | 設計対象 61 | 曖昧性 44 | 本方式 33 |
| ＧＡ 99 | 教育システム 61 | 利用者 44 | 知識ベースシステム 33 |
| 本手法 97 | 学習システム 60 | 背景知識 44 | 設計問題 32 |
| 事例ベース 95 | 本報告 59 | 制約充足 44 | 制約条件 32 |
| 遺伝的 90 | 人工知能 59 | 実験結果 44 | 熟練者 32 |
| 仮説推論 89 | エージェント間 59 | 高速化 44 | 自動化 32 |
| 対話システム 87 | 設計支援 58 | 概念設計 44 | 構成要素 32 |
| 類似性 85 | 言語処理 58 | 構文解析 43 | 処理システム 31 |
| 音声対話 83 | 帰納的 58 | 機械学習 43 | 有用性 30 |
| 設計者 78 | オブジェクト指向 58 | 設計知識 42 | 充足問題 30 |
| 最適化 77 | 定性的 57 | 法的 41 | 協調問題 30 |
| 意思決定 76 | 論理式 55 | 学習支援 41 | 理解状態 29 |
| モデル化 76 | | 情報処理 39 | 帰納推論 29 |
| 帰納学習 75 | | | |

Table 2

Top-100 terms when *tf-idf* is used.

| | | | |
|---|---|---|---|
| 学習者 496 | 音声認識 70 | オブジェクト指向 58 | 統合関係 20 |
| 問題解決 445 | 類似度 69 | 定式化 68 | 思考過程 36 |
| 知識ベース 229 | 対象モデル 72 | 概念設計 44 | 本報告 59 |
| 知的 243 | 設計者 78 | 故障仮説 23 | 機械翻訳 37 |
| ＧＡ 99 | 論理式 55 | 設計知識 42 | 一階 36 |
| 仮説推論 89 | ＣＡＩシステム 71 | 機能モデル 35 | 熟練者 32 |
| 支援システム 213 | 三面図 28 | 学習アルゴリズム 55 | 自動生成 46 |
| 決定木 62 | 法的 41 | 構造化 49 | 利用者 44 |
| 知識獲得 127 | 本手法 97 | ベース推論 55 | 知識コミュニティ 24 |
| 意思決定 76 | 相互作用 72 | 時間 63 | 物理現象 28 |
| 事例ベース 95 | 制約充足 44 | 言語処理 58 | 制御知識 25 |
| 知識表現 133 | 強化学習 50 | 背景知識 44 | 言語モデル 26 |
| 再利用 100 | エージェント間 59 | グループ学習 28 | 意味素 20 |
| 類似性 85 | 教材知識 47 | 構文解析 43 | 解決過程 39 |
| 遺伝的 90 | 本稿 594 | 設計過程 47 | 学習支援 41 |
| 本システム 142 | 動的 73 | 人工知能 59 | 統語 28 |
| 本研究 390 | モデル化 76 | 学習システム 60 | 情報処理 39 |
| 帰納学習 75 | 教育システム 61 | 自然言語 55 | 文字列 27 |
| 対話システム 87 | 定性推論 54 | 戦略知識 25 | 単一化 26 |
| 設計対象 61 | 故障診断 54 | 曖昧性 44 | 学習効果 26 |
| 音声対話 83 | 因果関係 54 | 高速化 44 | 多項式時間 34 |
| 本論文 420 | 推論システム 63 | 自動的に 68 | エージェント組織 23 |
| 最適化 77 | 定性的 57 | 充足問題 30 | 空間的 24 |
| 論理プログラム 69 | 設計支援 58 | 学習環境 46 | 最適解 36 |
| 有効性 166 | 帰納的 58 | 機械学習 43 | 協調問題 30 |

[**] The number of words is based on the output of our morphological analyzer, which contains segmentation errors.

[***] words such as “ 本論文(this paper)”

Table 3

Top-100 terms when MI is used.

| | | | |
|---|---|---|---|
| 連成 1 | 照合点 1 | 荷役ヤード 1 | 背腹 1 |
| 履修科目 1 | 小破断 1 | 科目届 1 | 背景色 1 |
| 落射 1 | 従属文 1 | 演劇経験者 1 | 東大工学部 1 |
| 有人観測所 1 | 秋葉三尺 1 | 渦巻ポンプ羽根車 1 | 鉄道台車 2 |
| 免疫ワードスポッティング 1 | 射照明 1 | 印加 1 | 提灯チョウチン 2 |
| 魔法陣 1 | 似顔絵師 1 | 意見 1 | 鳥類図鑑 2 |
| 付属テープ 1 | 資金使途 1 | 伊勢神宮 1 | 地区住民 1 |
| 瀬出し 1 | 残り体力 1 | ツルカメ算 1 | 大阪府高専 1 |
| 姫高原 1 | 三和銀行 1 | タイル取り 1 | 大阪大学溝口 1 |
| 費補助金 1 | 三尺坊 1 | サービス業務 1 | 相似異同 1 |
| 売上 1 | 埼玉県立 1 | オーストラリア国立 1 | 接続し実感 1 |
| 電動ウインチ 1 | 才児 1 | お札降り 1 | 製薬業 1 |
| 通謀虚偽 1 | 黒姫 1 | あき缶分別 1 | 数箇市町村 1 |
| 長野県黒 1 | 高級幹部 1 | 利用法 9 | 条通謀 1 |
| 中和滴定 1 | 公認会計士 1 | 有価証券 1 | 証券取引所 1 |
| 地中ライフライン 1 | 現実感 14 | 輸出立国 1 | 小売業 1 |
| 断冷却材 1 | 原油安 1 | 野外テント 1 | 受容器 2 |
| 大阪府立 1 | 県立久喜 1 | 模範文例 1 | 手書き帳 2 |
| 耐荷 1 | 空気ダンパ 1 | 鳴き真似 2 | 取り 1 |
| 川崎製鉄 1 | 京都大学西田 1 | 未定乗数 1 | 主査溝口 1 |
| 川喜田 1 | 久喜北 1 | 北陽 1 | 自走 1 |
| 水圧鉄管 1 | 喜田二郎 1 | 府立高専 1 | 指守 1 |
| 人称語尾 1 | 缶コーヒー 1 | 筆耕テクスト 1 | 索引付け 2 |
| 深部圧覚 1 | 冠婚葬祭 1 | 被災地 2 | 座標点 2 |
| 蒸留塔 1 | 改良策 1 | 発着信 1 | 混合音 1 |

Table 4

Top-100 terms when LLR is used.

| | | | |
|---|---|---|---|
| 本稿 594 | 動的 73 | 帰納的 58 | 熟練者 32 |
| 学習者 496 | 音声認識 70 | モデル化 76 | 有用性 30 |
| 問題解決 445 | 論理プログラム 69 | 言語処理 58 | 文脈自由 24 |
| 本論文 420 | 論理式 55 | 高速化 44 | 学習アルゴリズム 55 |
| 本研究 390 | 故障診断 54 | 機械翻訳 37 | 自己組織化 24 |
| 知的 243 | 時間 63 | 定性的 57 | 話し言葉 19 |
| 知識ベース 229 | 本システム 142 | 統語 28 | 試行錯誤 18 |
| 有効性 166 | 帰納学習 75 | 本報告 59 | 設計過程 47 |
| 支援システム 213 | 因果関係 54 | 最適解 36 | 設計支援 58 |
| 再利用 100 | 自然言語 55 | ベース推論 55 | 構成要素 32 |
| 相互作用 72 | 制約充足 44 | 教育システム 61 | 対象世界 34 |
| 意思決定 76 | 本手法 97 | 項目 27 | 評価値 17 |
| 決定木 62 | 曖昧性 44 | 設計対象 61 | 知識ベースシステム 33 |
| 事例ベース 95 | 対話システム 87 | ＧＡ 99 | 終端 16 |
| 知識獲得 127 | 設計者 78 | 文字列 27 | 可視化 28 |
| 人工知能 59 | 構文解析 43 | 自動生成 46 | 構造化 49 |
| 遺伝的 90 | 一階 36 | 入出力 25 | 極小限定 20 |
| オブジェクト指向 58 | エージェント間 59 | 背景知識 44 | 解決過程 39 |
| 仮説推論 89 | 定性推論 54 | 思考過程 36 | 制約条件 32 |
| 類似度 69 | 自動的に 68 | 教材知識 47 | 定理証明 22 |
| 最適化 77 | 対象モデル 72 | 物理現象 28 | 非線形 22 |
| 類似性 85 | 強化学習 50 | 実験結果 44 | 巡回セールスマン 15 |
| 音声対話 83 | 三面図 28 | 不適格 20 | 単一化 26 |
| 知識表現 133 | 創発 26 | 運転員 23 | 機械学習 43 |
| 定式化 68 | 多項式時間 34 | 述語論理 34 | |
| | | 現実感 14 | |

Table 5

Top-100 terms when *Rep*(·) is used.

| | | | |
|---|---|---|---|
| 学習者 496 | 仮説推論 89 | 方針決定 1 | 技術事業 1 |
| ＧＡ 99 | 事例ベース 95 | 独創的 1 | 技術開発 3 |
| 遺伝的 90 | 支援システム 213 | 調査システム 1 | 技術シーズ 1 |
| 先進工業 3 | 意思決定 76 | 第４報 1 | 基本業務 3 |
| 先行開発 2 | 機械翻訳 37 | 総合明確 1 | 基本基調 1 |
| 製品化 3 | 生産システム 3 | 先駆製品 1 | 基調的な 1 |
| 新製品 7 | 多項式時間 34 | 生存基盤 1 | 開発途上国指導 1 |
| 新技術 7 | 信念管理 7 | 水準高 1 | 開発目標 1 |
| 事業化 3 | 地球環境 15 | 資源無 1 | 開発市場 1 |
| 工業国 3 | 管理構造 7 | 仕様明確 1 | 開発基本 3 |
| 空洞化 2 | 文脈自由 24 | 昨年末 1 | 回復感 1 |
| 故障診断 54 | 製造業界 2 | 根本的 1 | マスコミ的 1 |
| 試行研究 4 | 問題解決 445 | 国民多 1 | ニーズ創造 1 |
| 論理プログラム 69 | 知的 243 | 国化 1 | シーズ創造 1 |
| 音声対話 83 | 一次変電所 16 | 高国民 1 | グローバル化 1 |
| 連想検査 4 | 再利用 100 | 構造改革 1 | 対象モデル 72 |
| 遠隔性 8 | 運転員 23 | 顧客ニーズ 1 | 参加者 23 |
| 音声認識 70 | 統合関係 20 | 原油安 1 | 相対位置 3 |
| エージェント組織 23 | 学習アルゴリズム 55 | 経済復興 1 | 自由文法 20 |
| 分散信念 6 | ＣＡＩシステム 71 | 景気回復 1 | 産業界 5 |
| エージェント間 59 | 因果関係 54 | 金融業界 1 | 最適解 36 |
| 本枠組み 9 | 類似度 69 | 業務明確 2 | 充足問題 30 |
| 最適化 77 | ベース推論 55 | 業務総合 1 | 設計過程 47 |
| 対話システム 87 | 輸出立国 1 | 教育水準 1 | 平均的 7 |
| | | 技術標準 1 | ファジィ理論 12 |
| | | 技術創造 1 | |
| | | 技術先行 1 | |

Table 6

Top-100 terms when LLR and *Rep*(·) are combined.

| | | | |
|---|---|---|---|
| 学習者 496 | 制約充足 44 | 現実感 14 | 見え方 11 |
| 問題解決 445 | 曖昧性 44 | 熟練者 32 | 実時間 21 |
| 知的 243 | 対話システム 87 | 文脈自由 24 | 直観主義 11 |
| 知識ベース 229 | 設計者 78 | 学習アルゴリズム 55 | 地球環境 15 |
| 有効性 166 | 構文解析 43 | 設計過程 47 | 人工生命 12 |
| 支援システム 213 | 一階 36 | 設計支援 58 | ブール代数 11 |
| 再利用 100 | エージェント間 59 | 評価値 17 | 適格文 14 |
| 相互作用 72 | 定性推論 54 | 制約条件 32 | 隣接 9 |
| 意思決定 76 | 対象モデル 72 | 巡回セールスマン 15 | 自由発話 16 |
| 決定木 62 | 強化学習 50 | 正例 24 | 決定支援 26 |
| 事例ベース 95 | 三面図 28 | 一次変電所 16 | 変電所事故 11 |
| 人工知能 59 | 創発 26 | 推論システム 63 | ストリーム分離 10 |
| 遺伝的 90 | 多項式時間 34 | 自由文法 20 | 学習支援 41 |
| 仮説推論 89 | 帰納的 58 | 充足問題 30 | マルコフ連鎖 9 |
| 類似度 69 | 言語処理 58 | 参加者 23 | セールスマン問題 15 |
| 最適化 77 | 機械翻訳 37 | 異常診断 22 | ゲーム木 12 |
| 音声対話 83 | 統語 28 | 学習環境 46 | ベース構築 23 |
| 動的 73 | 最適解 36 | 線形関数 20 | 刺激語 7 |
| 音声認識 70 | ベース推論 55 | 定性的な 36 | 解析法 22 |
| 論理プログラム 69 | 項目 27 | 各エージェント 27 | 近似解法 11 |
| 論理式 55 | 設計対象 61 | エージェント組織 23 | 統合関係 20 |
| 故障診断 54 | ＧＡ 99 | ＣＡＩシステム 71 | 空間内 15 |
| 時間 63 | 文字列 27 | 多重線形 15 | 認識率 14 |
| 因果関係 54 | 背景知識 44 | | 古典論理 11 |
| 自然言語 55 | 不適格 20 | | 所属性 12 |
| | 運転員 23 | | |
| | 述語論理 34 | | |

## 6. Conclusion

Our term extraction method uses grammatical filters and a novel measure for the representativeness of a term $W$. The basic idea of the measure is that the bias of the word distribution within the documents containing $W$ when compared with the background word distribution reflects the representativeness of $W$. The bias is measured by the normalized distance between the two distributions.

This measure has several advantages: (1) its definition is mathematically simple and clear, (2) it can naturally compare high-frequency terms with low-frequency terms, (3) the threshold value of being representative can be naturally defined, and (4) it can be applied to $n$-gram terms for any $n$.

Experiments show that this method is capable of picking out core terms with high precision. It eliminates highly frequent un-informative terms.

We plan to apply this measure to IR domain tasks such as construction of a stop-word list for indexing, and weighting terms in document-similarity calculation. A quantitative evaluation will also be conducted.

## Acknowledgement

The authors would like to thank Prof. Jun-ichi Tsujii (the University of Tokyo) for his valuable comments.

**References**

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., and 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. (1999) Measuring Representativeness of Terms, IPSJ Technical Report of SIGNL, 99- NL-133-17 (to appear, in Japanese)

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

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.

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

Nishioka, S., Niwa, Y., Iwayama, M., and Takano, A. (1997). *DualNAVI*: An information retrieval interface. Proc. of WISS'97, pp.43-48.

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.

Sakurai, H. and Hisamitsu, T. (1999) A Data Structure for Fast Lookup of Grammatically Connectable Word Pairs in Japanese Morphological Analysis, Proceedings of ICCPOL'99, Vol.II, pp.467-472

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.