

Searching for Explanatory Web Pages using Query Expansion

Manabu Tauchi Nigel Ward
School of Engineering, University of Tokyo
{manabu,nigel}@sanpo.t.u-tokyo.ac.jp

phone: +81-3-5841-6346;

fax: +81-3818-0835

<http://www.sanpo.t.u-tokyo.ac.jp/>

Mechano-Informatics, Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan

Summary

When one tries to use the Web as a dictionary or encyclopedia, entering some single term into a search engine, the highly-ranked pages in the result usually contain many irrelevant or useless sites. The problem is that single-term queries do not contain enough information to specify exactly which sort of pages the user wants. For the analogous problem in TREC, Buckley et al. (1996) have proposed query expansion, also known as pseudo-feedback or two-stage retrieval. In this method the top n documents returned by an initial retrieval are added to the query, which is then used for a second retrieval. This paper contributes, first, new normalization techniques for query expansion, and second, a new “local relevance density” metric which complements the vector product metric for computing similarity between an expanded query and a document. Both of these techniques are shown to be useful for single-term queries in Japanese over the World Wide Web.

Keywords

search engine, pseudo-feedback, local relevance density, terminology

1 Introduction

The Web (World Wide Web) contains massive amounts of information of every kind, constantly updated. We can access this information at home or from the office, for any purpose, and have come to rely on it more and more in place of traditional information sources. For example, when we want to investigate the meaning of a term, we frequently search for pages which describe or explain that term on the Web, instead of looking it up in a dictionary or encyclopedia.

However current search engines are not optimized for this use. Most models for information retrieval presuppose that the query contains sufficient information to meaningfully compute the similarity between the query and each document, but for single-term queries this is often not the

case. Thus the user typically has to refine his query, over several steps, adding terms to narrow the retrieval result to the sort of pages he wants.

In this paper, we present Perrie, a prototype ‘term search engine’. Perrie is designed to support the encyclopedic use of the Web, and its key techniques are automatic query expansion and evaluation of the relevance of web pages as providing explanation or discussion of specific terms. Our aim is similar to that of Sakurai’s [1] work, in that he was also interested in finding term explanations in Japanese on the Web, but he relied on syntactic pattern matching, and sought to find only pages containing definitions. In our study, the purpose of search is assumed to be not merely discovery of a term definition but more generally finding descriptions and explanations of that term.

2 Search Engine Basics

2.1 Web search tool

There are 2 kinds of web search tools, web directories represented by Yahoo and search engine such as Infoseek and Goo. In a web directory, web pages are categorized and indexed by human power. While they provide us more precise searching results, they cover many fewer web pages than search engine. Therefore, when we try to use the Web as a dictionary or encyclopedia, we mainly use search engines. But the number of search result pages found by search engine is so enormous and the precision of ranking pages is so low that it is difficult to find the page we want in them.

2.2 IR(Information Retrieval) model

Generally web search engines rely on the Vector Space Model and TF-IDF weighting. In the vector space model, all documents and queries are represented as vectors. There is one component in each vector for every distinct term that occurs in the document collection.

In many cases, the value of each component is calculated by TF-IDF weighting, which is the product of Term Frequency (TF) and Inverse Document Frequency (IDF). The following equation gives the weight of term T_j in document D_i .

$$W_{ij} = tf_{ij} \cdot \log \frac{N}{n}$$

W_{ij} :weight of Term T_j in Document D_i

tf_{ij} :frequency of Term T_j in Document D_i

N :number of Documents in collection

n :number of Documents where term T_j occurs at least once

In Vector Space Model, similarity is determined the inner product of the document vector calculated above and the query vector.

$$VSS_i = \vec{d}_i \cdot \vec{q}$$

VSS_i :Vector Space Similarity of page i

3 Query Expansion

When the query is a single term, we need to expand it by adding other words. Following Buckley et al. [2], we expand it by adding high frequency words from the collection of pages which contains the search word: these are assumed to be relevant and to give a more accurate query vector.

3.1 Buckley's Method

In pseudo-feedback (also called two-stage retrieval, Buckley et al. [2]), a short query is expanded as follows: After the first retrieval step, using the original short query, the top n documents returned by first retrieval are added to the query, and this is used for the second retrieval.

$$\vec{Q}_{expanded} = \vec{Q}_{original} + average\{\vec{d}_i | i \in (searchresult)\}$$

3.2 Creation of expanded query vector

Similarly, we assume that the average page vector of the page set containing the search term is generally usable as an enhanced or 'sharpened' equivalent of that search term.

$$\begin{aligned} & (DTf_1, DTf_2, DTf_3, \dots, DTf_n, \dots) \\ & = average\{\vec{d}_i | i \in (Searchresult)\} \end{aligned}$$

DTf_n :Domain term frequency (frequency of term n in the search result set)

Web pages have various lengths, so the vector d_i is normalized to have length 1 for each document. This is the first difference between our method and Buckley's.

Assuming then that more relevant words occur more frequently in the search result set, we define the degree of relevance to the search term as follows:

$$R_n = \max(\log(D_n/G_n), 0)$$

R_n :the Relevance of term n to the search term

D_n :the occurrence probability of term n in Domain web pages
($DTf_n / \sum DTf_i$)

G_n :the occurrence probability of term n in General web pages

In figure 1 showing R_n for a search on the Japanese term for “foot binding”, high relevant degree terms have something to do with search term, low ones have a little.

Term	R_{term}
纏足 (foot binding)	9.8
お婆さん (old woman)	7.3
奇習 (odd custom)	6.8
幼女 (infant girl)	6.4
宋 (T'ang Dynasty)	6.3
宦官 (eunuch)	6.2
足首 (ankle)	6.1
...	...
今 (now)	0.5
ハーブ (herb)	0.5

Figure 1: the example of R_n for the search term “foot binding”

This relevance is used as the weight of the word. Specifically, the expanded query vector \overrightarrow{ExQ} is defined to be $(R_1, R_2, \dots, R_n, \dots)$. To avoid problems in which few highly frequent words dominate the result, we take the logarithm of the frequency ratio. of the works. Also, to avoid problems of site-specific bias, for example to prevent a site which has the company name an every page from causing that name to get a high R score, a word which occurs in only pages from one site in assumed to be idiosyncratic to that site, and probably not relevant to the search word, and is given a weight of 0. These two refinements are also ways in which we improve on Buckley’s method to suit the web.

4 Web Page Evaluation

There are many factors which can be used to evaluate web page, including contents, link topology, update frequency, access count and

so on.[3] In this study, we restrict attention to content-based evaluation, as this factor is especially important when we search for pages containing meaningful information about a term.

We use the following two methods to evaluate Web page contents.

4.1 Vector Space Model

This model is explained in section 2.2. In our system similarity is determined the inner product of the normalized page vector and the expanded query vector, as follows:

$$VSS_i = \vec{d}_i \cdot \overrightarrow{ExQ}$$

VSS_i :Vector Space Similarity of page i

4.2 Local Relevance Density

The local density of relevant sentences is assumed to indicate the quality and quantity of description about search term a page.

Our rationale is that, of the various types of pages on the web (tables, lists, image archives), we are most likely to find term information in explanatory pages, and that these pages are comprised at least in part of clumps of related sentences, which include many words relevant to the query.

Thus the degree of relevance of a sentence is calculated as given by the equation below. However, when the same word occurs two or more times in a sentence, it is counted only once. A page is divided into sentences based on simple rules for sentence extraction considering HTML tags [4], followed by grammatical analysis using the Japanese morphological analysis system “Chasen” [5].

$$CR_k = \frac{\sum_i \sum_{j, i \neq j} R_i \cdot R_j}{N_k^2}$$

CR_k :the Co-occurrence-Relevance for the words of the sentence k

R_i :the Relevance of term i occurring in sentence k

N_k :the Number of terms in sentence k

The sentence which consists of relevant terms to the search term has high CR such as figure 2, and which consists of non-relevant has low such as figure 3.

纏足 は 中国 の 宋 の 時代 から
 9.8 3.1 6.3 1.8
 広まった 風習 です。
 5.0

(Foot binding is Chinese custom
 9.8 3.1 5.0
 in the age of T'ang.)
 1.8 6.3



$$\frac{9.8 \times 3.1 + 9.8 \times 6.3 + 9.8 \times 1.8 + \dots + 1.8 \times 5.0}{5 \times 5} = 21.8$$

Figure 2: the example of a sentence with high CR for search term “foot binding”

うち の ハーブ が 今 纏足 状態 です。
 0.8 0.5 0.5 9.8 0.5

(The herb in my house is in a state
 0.5 0.8 0.5
 like foot binding now.)
 9.8 0.5



$$\frac{0.8 \times 0.5 + 0.8 \times 0.5 + 0.8 \times 9.8 + \dots + 9.8 \times 0.5}{5 \times 5} = 2.0$$

Figure 3: the example of a sentence with low CR for search term “foot binding”

Since we assume that pages containing successive sentences which have a high Co-occurrence Relevance provide us much information, the following equation gives the Local Density of relevant

sentences on a page.

$$LD_i = \max\left\{\sum_k CR_k \times \max(10 - |x - k|, 0) \mid 0 \leq x \leq n_i\right\}$$

LD_i :Local Density of relevant sentences of page i

n :the number of sentence in the page

Figure 4 illustrates how the Co-occurrence Relevance varies across the sentences of a page: in this example, the page is crowded with relevant sentences near the middle. Figure 5 illustrates the window for computing the Local Density over each clump of nearby sentences. A 10 sentence half-width for the filter was found to give good results.

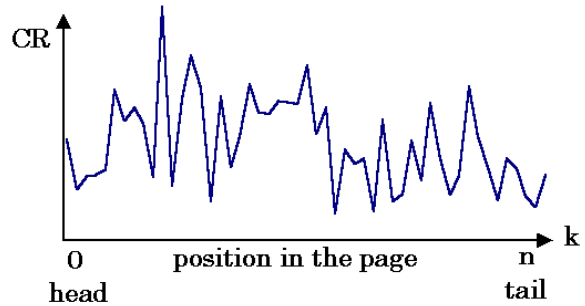


Figure 4: the example of CR distribution in a page

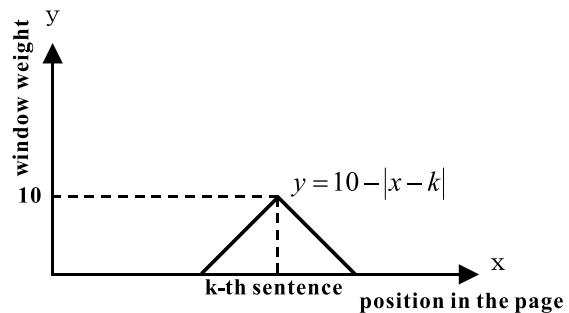


Figure 5: the window for measuring LD

5 Our Term Search Engine: Perrie

Figure 6 shows the process flow in Perrie. First, Perrie gets the top 100 pages searched and ranked by Infoseek (<http://www.infoseek.co.jp>). Next, using the morphological analysis system “Chasen”, nouns are extracted from each page, and the expanded query vector (§2.2) is created from these. The VSS (§3.1) and LD (§3.2) of each page are calculated from that vector, and the page is evaluated using both. Since the sort of page users probably want is highly relevant and has a dense explanatory section, we combine the two scores using the product.

$$P_i = LD_i \cdot VSS_i^\alpha$$

P_i :the evaluation score of page i

α :constant

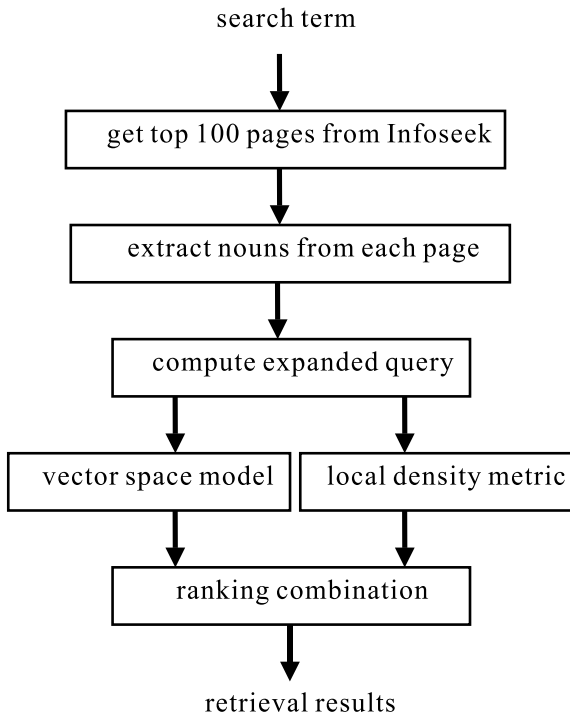


Figure 6: Overview of Processing in Perrie

6 Evaluation Experiment

6.1 Search Engine Evaluation Criterion

In Information Retrieval in general, some function of combining recall and precision is used for evaluating the quality of the document set returned. This method is based on the premise that all documents can be divided into two sets, those which are relevant and those which are not. However, this assumption fails for web pages, which are typically relevant or not to a degree. Moreover, unlike traditional information retrieval tasks, where users want all relevant documents, Web searchers typically are happy when they find any one page that has the information they want.

We therefore propose a new evaluation criterion. We have users score each page on a scale from one to seven, and the search engine’s ranking is considered better to the extent that its ranking correlates with the user’s scoring. Specifically we define the Raw Ranking Accuracy:

$$R.R.A. = \sum \frac{(\text{score of } i\text{-th ranked page})}{\log(i + 1)}$$

From this score, we normalize to obtain the Ranking Accuracy as follows:

$$RA = \frac{R.R.A. - \text{Expected } R.R.A.}{\text{Ideal } R.R.A. - \text{Expected } R.R.A.}$$

The ideal R.R.A. here is the evaluation score if the pages are ordered perfectly in accordance with the user’s scores, and the expected R.R.A. is the expected value of for a search engine that ranks the pages of the set at random. Normalization sets the Ranking Accuracy of a random engine at 0, and of an ideal engine at 1.

6.2 Experiment Method

The subjects were given the following task. “Please do term search from the Internet in order to acquire the meaning or the some information about a word. Please think of some word about which you want to learn more.” All subjects chose

Japanese words and most gave no scores to non-Japanese pages. The word chosen by the subject was fed to our baseline Infoseek, and the top 40 search result pages were scored by the subject from 1 (not at all related/ not at all helpful) to 7 (very satisfying). We asked users to base their scoring on the content of the page itself, not considering the value of pages linked to from that page.

6.3 Result

Subjects came up with the following 19 terms:

- a. IT 革命 (IT revolution)
- b. 出師表 (Chinese historical address)
- c. ADHD (Attention Deficit Hyperactivity Disorder)
- d. 関羽 (name of historical person)
- e. 屈原 (name of historical person)
- f. 直木三十五 (founder of Naoki prize)
- g. ユーゴスラビア (Yugoslavia)
- h. マイライン (default telephone provider)
- i. 行為障害 (conduct disorder)
- j. オギノ式 (rhythm method)
- k. 筋ジストロフィー (muscular dystrophy)
- l. ニューディール政策 (New Deal)
- m. ラマーズ法 (Lamaze method)
- n. 纏足 (foot binding)
- o. 金庫株 (new stock system in Japan)
- p. ワルサーP38 (make of revolver)
- q. スウィングバイ (swing by)
- r. ハプスブルグ家 (House of Hapsburg)
- s. 宮部みゆき (Japanese famous author)

For each of these 19 terms, the 40 web pages were ranked by four page ranking methods. For each method, the ranking accuracy was computed using the top-20 ranked pages, based on the observation that most users will not examine more than about 20 hits. The four methods were:

- A. Ranking pages using the vector space model with conventional query expansion, namely pseudo-feedback.
- B. Ranking pages using the vector space model and the expanded query as computed by Perrie.
- C. Ranking pages using the local density metric and the expanded query as computed by Perrie.
- D. Perrie, combining B and C, with α set to 1.0.

The ranking accuracy of 19 terms by Infoseek and methods of A, B, C, and D is shown in Table 1, and the average of that is shown Figure 7. Statistically significant differences are observed for all pairings except BC.

	Info	A	B	C	D
a.	.39	.39	.41	.65	.52
b.	.25	.24	.78	.59	.78
c.	-.21	.48	.60	.73	.77
d.	-.17	.66	.84	.85	.86
e.	.09	.39	.30	.45	.48
f.	.30	.48	.27	.76	.80
g.	.17	.05	.26	.49	.44
h.	-.36	.32	.37	.48	.56
i.	.07	.23	.48	.39	.41
j.	.25	.39	.60	.68	.57
k.	.43	.64	.47	.37	.41
l.	-.16	.22	.44	.49	.63
m.	.05	.50	.67	.77	.78
n.	.16	.80	.75	.11	.71
o.	.19	-.10	.70	.61	.84
p.	.10	.74	.81	.77	.89
q.	.15	.38	.54	.45	.60
r.	-.32	.60	.64	.35	.60
s.	.15	.50	.62	.70	.74

Table 1: Ranking Accuracy of terms by each method

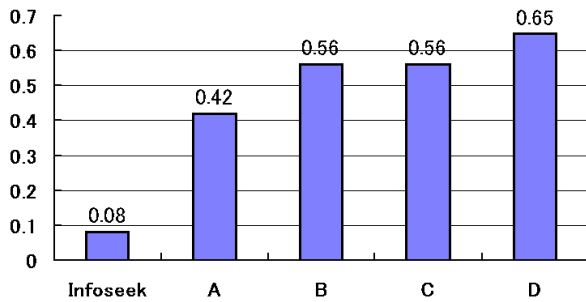


Figure 7: Average Ranking Accuracy

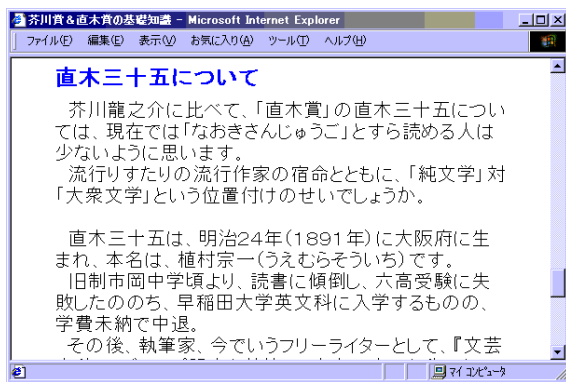


Figure 8: The 1st ranked page for search term f.

7 Discussion

Figure 7 shows that for average ranking accuracy, an average difference of 0.14 points was observed between the conventional method, pseudo-feedback (A), and our method for computing the expanded query vector which normalizes with respect to baseline word frequency across the web (B).

Next, comparing among methods of B, C, and D, which all use our extended query vector, no significant difference was seen between method B, using the vector space model and C, using local density. However in method D (Perrie), using both, the ranking accuracy rises 0.09 point compared with methods B and C on average, and this is statistically significant. This rise is considered to result from the fact that the combination of the two evaluation models is effective at penalizing low quality pages.

Analysis of cases where valuable pages were given low scores by Perrie reveals 4 main causes:

- The search result set consisted of two or more clusters, due to a term with two senses or appearing in web pages of two unrelated types. For example, when subject chose the term g. “Yugoslavia”, themes of almost all searched pages were categorized into two: soccer and the situation of the nation. Ranked top pages consisted of these two kind of pages. But the subject wanted only information about the latter. So the half of the top pages were given poor scores.
- The relevant degree of a term was too high due to a spurious correlation. Figure 9 shows the 5 highest relevant degree terms to the search term q. “swing by”. While all terms except for “Nostradamus” are somehow relevant to “swing by”, “Nostradamus” is not. This phenomenon results from a spurious correlation. In this case, “Cassini” is a famous spacecraft which changed orbit by the navigation technique called “swing by”, so these two terms coincided. And some documents say that the catastrophe in August, 1999 prophesied by “Nostradamus” results from “Cassini” impacting on the earth. Here, “Nostradamus” and “Cassini” coincided, too. Therefore “Nostradamus” and “swing by” often coincided though “Nostradamus” does not correlate with “swing by” directly.

Term	R_{term}
カッシーニ (Cassini)	8.9
ノストラダムス (Nostradamus)	7.7
スイング (swing)	7.4
公転 (revolution)	7.3
火星 (Mars)	7.3

Figure 9: the Highest relevant degree term in “swing by”

- Vision information, such as images, were un-

modeled factors in a subject's page scoring.

- One subject scored an English language page highly but of course the expanded query matched that page poorly.

Compared with Infoseek, the search engine Perrie obtained a higher ranking accuracy in 18 of 19 words, showing an increase of 0.57 points on average. Moreover, at least one of the pages scored highest by the subject was found among top 3 ranked pages for 14 of the 19 words; an example is seen in Figure 8.

8 Conclusion

In searching for explanatory Web pages, the expanded query vector and the local density of relevant sentences proposed by this study were proved to be effective for web page evaluation. Moreover, it can be said that a search engine effective for encyclopedia or dictionary-use of the WWW was feasible by these methods.

References

- [1] Yuu Sakurai, Satoshi Satoh. Using the World Wide Web for Terminology Search. Information Processing Society of Japan, Workshop Notes. 2000-NL-137-4, 2000.
- [2] Chris Buckley, Amit Singhal, Mandar Mitra, and Gerard Salton. "New Retrieval Approaches using SMART:TREC4". *The Fourth Text REtrieval Conference (TREC-4)*. In D.K.Harman, 1996.
- [3] Shun-ichi Fukushima. The Future of WWW Search Engines and Evaluation Methods. *Information Processing Society of Japan Magazine*. Vol.41, No.8, pp913-916, 2000.
- [4] Yoji Kiyota, Sadao Kurohashi. Gisting and KWIC-index Generation from Web Texts. Information Processing Society of Japan. Workshop Notes. 2000-NL-137-5, 2000.
- [5] Yuji Matsumoto, Akira Kitauchi, Yoshitaka Hirano, Tatu Yamashita, Masayuki Asahara. Users Manual for Chasen version 2.0; A Morphological Analyzer for Japanese. JAIST-Nara technical report. 1999.