Improving Web Search by Query Expansion with a Small Number of Terms

Tomonari Masada
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
masada@nii.ac.jp

Teruhito Kanazawa
KYA Group Corporation
5-29-7 Koishikawa, Bunkyo-ku, Tokyo 112-0002, Japan
tkana@kyagroup.com

Atsuhiro Takasu Jun Adachi
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
takasu, adachi@nii.ac.jp

Abstract

This paper presents comprehensive experimental results on query expansion (QE) with a small number of terms. Our retrieval process uses a term weighting scheme based on a probabilistic model. The terms of large Robertson's selection values are used for QE. In most QE experiments, hundreds of terms are added to an original query. However, this style of QE substantially increases search response time. In this paper, we provide experimental results attained by using at most 20 terms for expansion. The best average precision is nearly 8% increase over the baseline case, i.e., the case where we do not use QE. This improvement is obtained when only ten terms are added to an original query. Our results show that we can achieve a fairly good improvement even with a small number of expansion terms. The results also show how we should adjust parameter values to improve the quality of seach systems using QE.

Keywords: *WWW, Information Retrieval, Query Expansion.*

1 Introduction

This paper presents comprehensive experimental results on query expansion (QE) with a small number of terms. Our retrieval process uses a term weighting scheme based on a probabilistic model after a slight modification reported in [2]. Then, the terms having large Robertson's selection values [7] [6] are used for QE. We limit the number of expansion terms to at most 20. In most QE experiments, hundreds of terms are added to an original query[10][12]. However, this style of QE substantially increases search response time and introduces a complication with respect to the actual system implementation, e.g. parallelized or distributed processings. In this paper, we present experimental results attained by using at most 20 terms for QE. The best average precision is nearly 8% increase over the baseline case, i.e., the case where we do not use QE. This improvement is obtained when only ten terms are added to an original query. Moreover, expansion terms are selected only from the top five documents of an initial retrieval. Our results show that we can achieve a fairly good improvement even with a small number of expansion terms. Therefore, it can be concluded that this kind of light-weight QE technique is a realistic choice for an efficient and scalable Web search. However, to improve search quality with light-weight QE, we should adjust system parameters appropriately. The parameters to be adjusted are:

1. Control parameters appearing in the term weighting formula,
2. Number of the top-ranked documents from where expansion terms are extracted, and
3. Mixture ratio multiplied to the weights of expansion terms when we take a summation of all term weights for computing a document score.

According to our experimental results, we can improve Web search fairly well with a wide range of values of

the second parameter, i.e., the number of top-ranked documents from where expansion terms are extracted. However, the first and the third parameters heavily affect the search quality. Therefore, our findings must provide a practical and informative guide in choosing parameter values appropriately to design an efficient light-weight QE system.

2 Information retrieval model

We adopt a probabilistic model and use a term weighting formula as presented in [9], which is a variation of BM25 [8]. However, we modify the formula as reported in [2]. That is, we replace the so-called Robertson-Sparck Jones weight with the logarithm of IDF appearing in a traditional term weighting formula, TF-IDF [1]. This replacement actually improves the quality of our Web search. Futher, in our experiment, no terms appear more than once in the query. When we do not use free text queries, this is an ordinary case. Therefore, we omit the factor of query term frequency from our weighting formula. As a result, we use the following as the term weight $w(t)$ of a term t appearing in a document d:

$$w(t) \equiv \log\Big(\frac{N}{df}\Big) \cdot \frac{(k+1)\cdot tf}{k\{(1-\alpha)+\alpha\frac{dl}{avdl}\}+tf} \quad (1)$$

where

N: Number of documents in the collection

df: Document frequency of t

tf: Term frequency of t in d

dl: Document length of d

$avdl$: Average document length

k, α: Control parameters.

To avoid overtuning control parameters, we simply set α to 0.5, because this value can attain fairly good retrieval results. As for k, we test the following four values: 10, 20, 50, and 100.

When we compute the score of a document d with respect to an initial query q, i.e., a query before expansion, we take a summation of the term weights of all terms included both in q and d. On the other hand, when we compute the score of a document with respect to an expanded query q', the weights of terms used for expansion are multiplied by a constant $\mu < 1$. In this paper, we call this constant μ *mixture ratio*. Consequently, the score of a document d with respect to an expanded query q' is computed as follows:

$$Score(d) \equiv \sum_{t\in q\cap d} w(t) + \mu \cdot \sum_{t\in(q'\setminus q)\cap d} w(t). \quad (2)$$

3 Term selection method

Our term selection method is a conventional one, i.e., a method using Robertson's selection value (RSV) [6]. The selection weight of a term t is defined as follows:

$$RSV_t \equiv \Big(\frac{rdf}{R} - \frac{df}{N}\Big) \cdot \Big\{\beta \log\frac{N}{df} + (1-\beta)\log\frac{\frac{rdf+0.5}{R-rdf+0.5}}{\frac{df-rdf+0.5}{N-df-R+rdf+0.5}}\Big\} \quad (3)$$

where

R: Number of relevant documents

rdf: Number of relevant documents including t

β: Control parameter.

Here too, we set $\beta = 0.5$ to avoid overtuning. As for R, we test the following values: 5, 10, and 20. The terms appearing in the top R documents of an initial retrieval are sorted by their selection weights. Then, top-ranked terms are used for QE. Let T be the number of expansion terms. In our experiment, T takes either of the following values: 5, 10, or 20. In addition, we test 15 values for μ in Eq. 2, ranging from 0.005 to 0.2, i.e., 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, and 0.2. As a result, we test $4 \times 3 \times 3 \times 15 = 540$ combinations of values of four paremeters, k, R, T, and μ.

4 Experiment settings

We use a document set prepared for NTCIR-5 WEB Query Term Expansion Subtask [11]. As all documents are written in Japanese, we decompose each document into a multiset of terms by using a morphological analyzer MeCab[1]. We also use an open dictionary ipadic-2.5.1[2] without any modification. Initial retrievals are executed with 35 queries prepared for the task. Then, we regard the top R documents as relevant documents for each query. As mentioned above, we test the values 5, 10, and 20 for R. From the R relevant documents, we select the terms for QE by using RSV (See Eq. 3) and expand each original query by the top-ranked terms. The number of terms used for QE, denoted by T, is either 5, 10, or 20. The top 10 terms are listed in Table 1 and 2 for each of the 35 queries and for all three settings of the parameter R. In a retrieval with an expanded query, document scores are computed with Eqs. 1 and 2. We evaluate all retrieval results by comparing them with the relevant document set provided by the organizer. This paper presents the quality of each retrieval result with the average precision and the R-precision.

[1] http://chasen.org/~taku/software/mecab/

[2] http://chasen.naist.jp/stable/ipadic/

5 Results of Experiments

Table 3 includes all average precisions as the ratio to the baseline value 0.1620. The leftmost column shows the experiment settings. k stands for the control parameter appearing in Eq. 1, R for the number of top-ranked documents from where we extract expansion terms, and T for the number of expansion terms. For example, the row with the leftmost entry "10/5/20" includes the average precisions attained when we set $k = 10$, $R = 5$, and $T = 20$. Each row includes 15 average precisions obtained with 15 different values of mixture ratio μ ranging from 0.005 to 0.2. For example, the column with the uppermost entry "0.08" includes the average precisions when we multiply 0.08 to the weights of expansion terms as shown in Eq. 2. On the other hand, Table 4 includes all R-precisions as the ratio to the baseline value 0.2066. This table has the same structure as Table 3. In both tables, each italicized number is the largest one for a fixed k. The largest number in each table is boldfaced. We can obtain the best average precision when $k = 20$, $R = 5$, $T = 10$, and the mixture ratio μ is 0.14. We can also obtain the best R-precision when $k = 10$, $R = 10$, $T = 5$, and $\mu = 0.10$. The average precisions and the R-precisions for $k = 20$, $R = 5$, and $T = 10$, are presented in Figure 2, Figure 3, and Figure 4, respectively. The left panel shows the average precisions, and the right panel shows the R-precisions. Although the average precisions and the R-precisions for $R = 5$ show less dispersion than those for $k = 20$ and those for $T = 10$, it can be concluded that a careful tuning of each parameter is required to obtain a large improvement. Table 3 and Table 4 must provide a practical and informative guide in choosing parameter values appropriately.

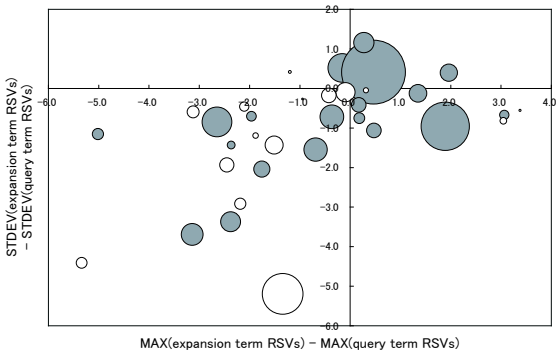

Figure 1. Web search improvements characterized by the correlations between query term RSVs and expansion term RSVs (See also Table 5.)

6 Analysis

In this section, we analyze how the average precisions of 35 queries are improved by QE. In our analysis, we set $k = 20$, $R = 5$, $T = 10$, and $\mu = 0.14$, because this combination of parameter values gives the best average precision among those obtained in our experiments. We characterize the average precision improvement of each of 35 queries by the correlation of the following two values:

- Difference between the maximum of expansion term RSVs and that of query term RSVs, and
- Difference between the standard deviation of expansion term RSVs and that of query term RSVs.

When the former value is positive, there is at least one expansion term whose RSV is larger than any of the query term RSVs. Therefore, we can expect QE to contribute to the improvement of Web search quality. For example, the query of ID 0098 consists of the following terms: "世界遺産 (world heritage)" and "日本 (Japan)." After a morphological analysis, the term "世界遺産" is decomposed into two terms: "世界 (world)" and "遺産 (heritage)." Consequently, this query consists of three query terms. For this query, we have a term "ユネスコ (UNESCO)" as the expansion term having the largest RSV. The RSV of "UNESCO" is equal to 8.7684 and is larger than 6.8757, which is the RSV of "heritage". The term "heritage" has the maximum RSV among the three query terms. The RSV of "world" is 3.7056, and that of "Japan" is 2.9436. As for this query, QE technique substantially improves the average precision from 0.0729 to 0.1863.

On the other hand, when the latter value, i.e., the difference between the standard deviation of expansion term RSVs and that of query term RSVs, is small, the query term RSVs show a larger divergence than the expansion term RSVs. The query terms showing a large divergence in their RSVs may not be able to create a good topical unity in an initial retrieval result. Therefore, if the expansion terms extracted from such a retrieval result show a small divergence in their RSVs, these expansion terms may give a search result focusing on a topic apart from that expressed in the query terms. Consequently, it is desirable that the expansion term RSVs also show a fairly large divergence when the query term RSVs show a large divergence. Our latter value captures this intuition. For example, the query of ID 0097 consists of two query terms: "アイメイク (eye makeup)" and "やり方 (method)." The RSVs of these two terms are 11.2245 and 0.9881, respectively, and show a large divergence. The expansion terms for this query include "目元 (the expression of the eyes)", "まぶた (eyelid)", and "色 (color)". Since any of these terms can appear in a context unrelated to the eye makeup, they may cause a topic drift.

In fact, the average precision of this query decreases from 0.1340 to 0.0552 by using QE.

Figure 1 characterizes the Web search improvements of all 35 queries by the correlations between the two values shown above. The horizontal axis represents the former value, i.e, the difference between the maximum of expansion term RSVs and that of query term RSVs. The vertical axis represents the latter, i.e., the difference between the standard deviation of expansion term RSVs and that of query term RSVs. Each gray-filled circle stands for a query with respect to which we can improve the average precision by QE. Each white-filled circle stands for a query with respect to which QE makes the average precision decrease. The area of each circle shows the amount of increase (for a gray-filled circle) or decrease (for a white-filled circle) in average precision. It should be noted that we have many gray-filled circles on the upper right side of the figure. All data are presented in Table 5.

7 Link-based clustering

In [4], we proposed a link-based Web page clustering method. By using this method, we can modify document scores and improve Web search. As mentioned in Section 5, our experiments provide the best average precision when $k = 20$, $R = 5$, $T = 10$, and $\mu = 0.14$. We fix these parameter values and modify document scores by using a result of Web page clustering proposed in [4]. This clustering result is obtained by executing a *Cyclic clustering* with $\tau = 35$. τ is a parameter, called *threshold paramter*, and controls the granularity of clusters. We refer readers to [4] and [5] for the details of Cyclic clustering of Web pages. The procedure for document score modification is as follows:

1. For each cluster C of Web pages, we compute the maximum score among the pages in C. Denote the maximum score as $MaxScore(C)$. This score can be regarded as the score of a cluster C.

2. For each cluster C, we modify the score $Score(d)$ of a Web page d in C as follows:

   $$Score'(d) \equiv (1-\gamma)Score(d)+\gamma MaxScore(C)$$

   where γ is a paremeter. The definition of $Score(d)$ appears in Eq. 2.

As described above, we obtain a new document score as a linear mixture of an original document score and a cluster score. This procedure is a simplified version of that in [3]. When we set $\gamma = 0.4$, we can obtain an average precision 0.1810, which is 11.7% increase from the baseline, and an R-precision 0.2236, 8.2% increase from the baseline. This result proves that our Cyclic clustering can improve Web search. However, this techinique has no direct relation with QE technique. Therefore, we add the above result as an additional one with respect to QE task.

8 Conclusions

As is well known, QE technique introduces a lot of parameters to be tuned into an information retrieval system. Therefore, we conducted a long series of experiments and presented all the results of our experiments in this paper. The results show that we can obtain a fairly good improvement of Web search even with a small number of expansion terms. However, a careful parameter tuning is required. Our comprehensive experimental results must provide a cue to effectively manage a search system using QE with a small number of terms.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *SIGIR'04*, pages 49–56, 2004.

[3] T. Kanazawa, A. Takasu, and J. Adachi. A relevance-based superimposition model for effective information retrieval. *IEICE Trans. Inf. & Syst.*, E83-D(12):2152–2160, 2000.

[4] T. Masada, A. Takasu, and J. Adachi. Link-based clustering for finding subrelevant Web pages. In *Proceedings of the Third International Workshop on Web Document Analysis*, 2005.

[5] T. Masada, A. Takasu, and J. Adachi. Improving Web search performance with hyperlink information. *IPSJ Transactions on Databases*, 46(SIG8):48–59, 2005 (in Japanese).

[6] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359 – 364, 1991.

[7] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[8] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC-3*, pages 109–126, 1994.

[9] A. Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–43, 2001.

[10] M. Toyoda, M. Kitsuregawa, H. Mano, H. Itoh, and Y. Ogawa. University of tokyo/RICOH at NTCIR-3 Web retrieval task. In *Proceedings of the Third NTCIR Workshop*, 2002.

[11] M. Yoshioka. Overview of the NTCIR-5 WEB query term expansion subtask. In *Proceedings of the Fifth NTCIR Workshop*, 2005.

[12] M. Yoshioka and M. Haraguchi. Study on the combination of probabilistic and boolean IR models for WWW documents retrieval. In *Proceedings of the Fourth NTCIR Workshop*, 2004.

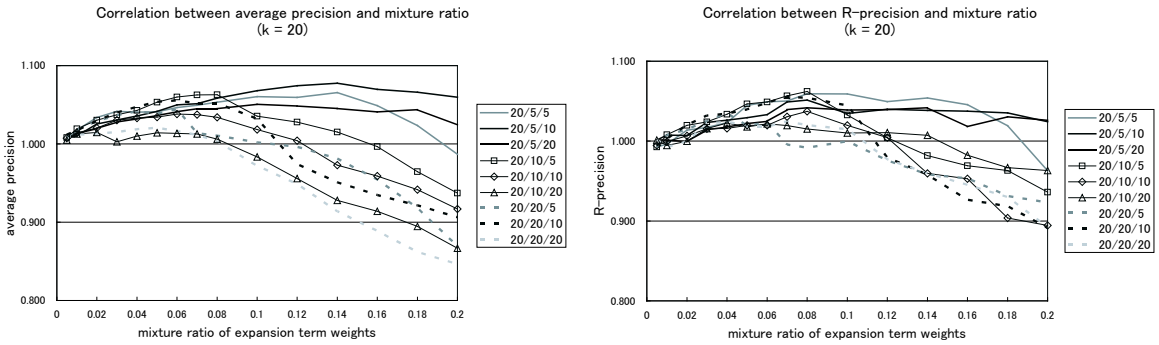

Figure 2. Evaluation results for $k = 20$.

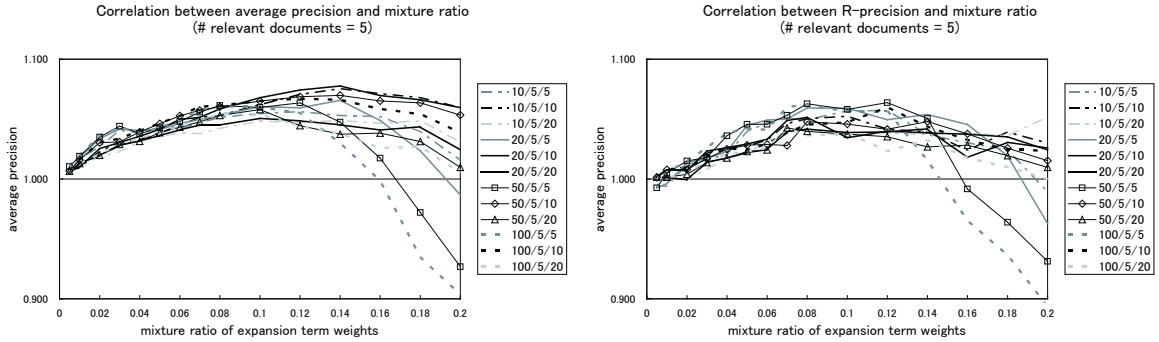

Figure 3. Evaluation results for $R = 5$.

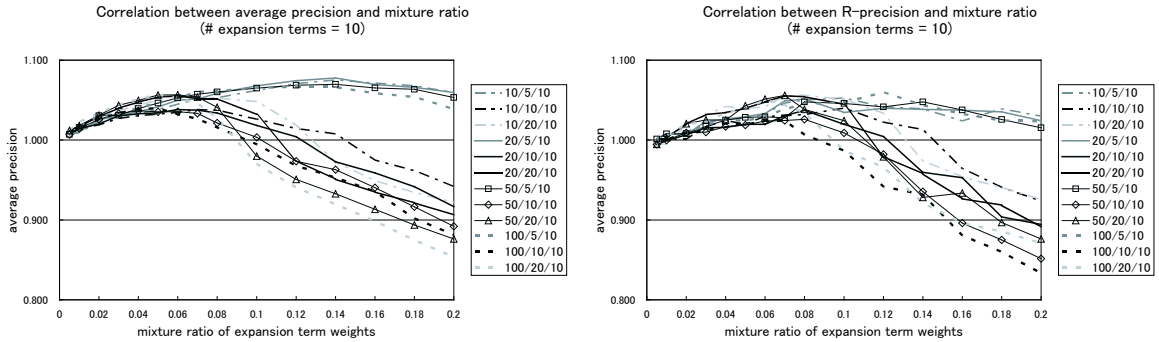

Figure 4. Evaluation results for $T = 10$.

Table 1. Top 10 expansion terms (query IDs: from 0001 to 0076)

qid	R	top 10 terms
0001	5	ボール, 選手, ポジション, ゴール, フォワード, ディフェンス, センターフォワード, 戦術, オフサイドトラップ, 味方
0001	10	ディフェンス, パス, 反則, ボール, センターフォワード, フォワード, ゴール, トラップ, スポーツドリンク, ストレッチング
0001	20	パス, ゴール, ボール, ディフェンス, 反則, チーム, プレー, 戦術, 選手, キック
0003	5	プトレマイオス, 天動説, ガリレオ, 天文学, 太陽, 法則, 教会,1473,1543, ブラーエ
0003	10	太陽, 天動説, ガリレオ, プトレマイオス, 木星, 天文学, 唱える, フィレンツェ, 法則, 観測
0003	20	天動説, ガリレオ, 天文学, プトレマイオス, 太陽, ケプラー, 観測, 地球, 天体, ガリ
0004	5	ベツレヘム, 元子, 当事者, 西岸, 和平, 政経, 神前, 彼ら, 期生, 中東
0004	10	西岸, 和平, ユダヤ, 入植, エルサレム, ベツレヘム, ガザ,Jerusalem, アラブ, 中東
0004	20	アラファト, ガザ, 和平, 西岸, ユダヤ, 入植, エルサレム,Jerusalem, アラブ,israelwire
0006	5	Slew, 勝馬, せん馬, 斤量,Gulch, 牡馬, アタマ,Nureyev,Quest, 牝馬
0006	10	せん馬, 勝馬, 斤量, 牡馬, アタマ, 牝馬, クビ, 騎手,CLICK,HERE
0006	20	牡馬, せん馬, 勝馬, 斤量, 牝馬, 馬, クビ, 騎手, アタマ,Melancon
0019	5	美術, 彫刻, 絵画, 鑑賞, 宗教, ルネサンス, 史, キージ, アンモニヤ, ポピーオイル
0019	10	世紀, 美術, 彫刻, 絵画, アンモニヤ, ポピーオイル, カバコフ, 舞踏, ルネサンス, 意味
0019	20	絵画, 美術, ルネッサンス, 世紀, 歴史, イタリア, 表現, 史, 意味, 時代
0021	5	値うち, 柳美里,mediatv,JUSTICE,justice, 批評, 日本人, 目玉, 文芸,portrait
0021	10	柳美里,mediatv,JUSTICE,justice, 値うち, 冊, 批評, 日本人, 書く, 文芸
0021	20	mediatv,JUSTICE, 柳美里,justice,portrait,jump,banner,Part, 江藤,Profile
0022	5	グレイ, アール, 紅茶, ティー, ブレンド, フルーツ, ジャスミン, ラプサンスーチョン, サーティーフーズ, ミント
0022	10	紅茶, ティー, ブレンド, グレイ, 香り, アール, アイスティー, フレーバー, ローズ, レモン
0022	20	紅茶, ティー, ブレンド, 香り, アイスティー, グレイ, ミルク, 甘い, フルーツ, アール
0023	5	00010411,00001541,00022390, 酒客, ダイエット, みかの,Darlin, 栄養, 酒宴, パケ
0023	10	00010411,00001541,00022390, 栄養, 酒客, みかの,Darlin, ダイエット, 酒宴, パケ
0023	20	栄養,00010411,00001541,00022390, 酒客, みかの,Darlin, ダイエット, 健康, 豆
0028	5	所得,mygate, 累進税, 控除, 新倉, 課税, 翌年, 年末, 額, 道府県
0028	10	額, 所得, 給与, 控除, お金, 支払う, の, かかる, 金額, 万
0028	20	所得, 給与, 税, 額, 金額, 天引き, 控除, 明細, 毎月, お金
0029	5	QR,VAIO,SR,505, モデル,PCG,sony,RX, デスク,MX
0029	10	RX,VAIO,LX,PCV,MX,XP, ビデオカメラ, モデル, ワンズ,sony
0029	20	VAIO,Hundycam, ビデオカメラ, グランドベガ, アイボ, ワンズ,pcv,RX,trv, ハンディカム
0034	5	薄切り, 味, 輪切り, きゅうり, みじん切り, 包丁, 切る, 拍子木, ささがき, 乱切り
0034	10	包丁, 拍子木, 切る, 薄切り, 味, みじん切り, いちょう, 輪切り, 半月, ささがき
0034	20	包丁, 切る, 味, 煮る, 輪切り, 拍子木, 野菜, 薄切り, 短冊, 半月
0044	5	語学, プログラム, ビジネス, 講師, 名古屋,WISH, 陣, 管理, ランゲージ, コース
0044	10	語学, ホームステイ, プログラム, ビジネス, 学校, コース, 力, 海外, 留学,WISH
0044	20	ホームステイ, プログラム, 海外, ビジネス, 語学, 力, コース, 留学, スキル, 派遣
0045	5	食べる, ソース, アンジュ, マカロニ, スパゲッティ, ミート, サンタ, リガーテ, 初夏, ブカティーニ
0045	10	リガーテ, ソース, 食べる, マカロニ, スパゲッティ, トマト, ラザニエ, シチリア, タリアッテレ, ゆでる
0045	20	マカロニ, ソース, トマト, リガーテ, スパゲッティ, イタリア, 食べる, チーズ, ペン, イタリアン
0055	5	シーレ, ウィーン, 観する, ベルベデーレ, 美術, そばだてる, 落ち合う, トラム, グスタフ, 末
0055	10	シーレ, ウィーン, 美術, グスタフ, 世紀, 末, ロマン, ココシュカ, 観する, 彫刻
0055	20	シーレ, ウィーン, 美術, グスタフ, 世紀, 末, 画家, 絵画, 作品, コレクション
0058	5	思想, 人間, 史, 顕現, 思惟, 認識, 様態, 論述, 弘文, 主義
0058	10	解釈, 思想, 人間, 形而上学, 主義, 西洋, 史, ハイデッガー, 学, 根本
0058	20	西洋, 主義, 学, 思想, 形而上学, 人間, 倫理, 現代, 認識, 的
0061	5	派遣, 全域, 修学, 進学会, 教材, 無料, アルバイト, 黒須, 進学, 塾
0061	10	派遣, 塾, 指導, 無料, トライ, 全国, 受験, カテキョ, 全域, 生徒
0061	20	派遣, 指導, 塾, トライ, 生徒, アルバイト, テューター, 無料, 受験, 全国
0062	5	スギ, スカイナー, 慈恵, アレルギー, 鼻炎, 鼻かぜ, 鼻水, 耳鼻科, カバノキ, 療法
0062	10	アレルギー, スギ, 飛散, スカイナー, 鼻水, 治療, 療法, 鼻かぜ, 慈恵, 薬
0062	20	アレルギー, 治療, スギ, 飛散, 鼻, 鼻水, 症状, 薬, マスク, くしゃみ
0063	5	骨折, 裂, 靱帯, 関節, 断, 損傷, 腱, 復帰, 骨, 早期
0063	10	骨折, 関節, 裂, 損傷, 捻挫, 障害, 足, 断, 復帰, アキレス腱
0063	20	捻挫, 骨折, 関節, 膝, 痛み, テーピング, 外傷, 裂, 足, 障害
0065	5	態度, 気づく, 医学, 気分, 身体, 一如, 精神, 受診, 密接, 私
0065	10	気分, 態度, 精神, 症状, 性格, 行動, 一如, 二元論, 症, 医学
0065	20	精神, 身体, 性格, 呼吸, 一如, 気分, 行動, 状態, 治療, それ
0068	5	稽古, 武道, 段, 段位, 剣, 寒げいこ, 初段, 武道館, げいこ, 高段
0068	10	武道, 稽古, 段, 剣, 初段, 修行, 受審, 武道館, 合格, 大会
0068	20	稽古, 段, 剣, 武道, 武道館, 初段, 大会, 竹刀, 体育館, 道場
0070	5	喫煙, 原告,maedad, 未成年, 禁煙, 遺族, 嫌煙,tachibana, エレテック,Claims
0070	10	喫煙, ニコチン, 禁煙, 嫌煙, 原告, 分煙,maedad, たばこ, 吸う, 受動
0070	20	喫煙, ニコチン, 禁煙, 受動, 吸う, 煙, 健康, 分煙, 嫌煙, たばこ
0071	5	Population, 総務庁, 泰代, 宮城学院女子大学,Projections, 推計, 所内, エイジング, 東北学院, 出生
0071	10	Population, 統計, 出生, 学会, 総務庁, 調査,1998, 推計, 所内, 少子
0071	20	Population, 推計, 出生, 総務庁, 統計, 動態, 少子化,Social, 保障, 学会
0073	5	ブルズ,MJRetirement,NBA, ウィザーズ, 引退, チーム, 復帰, 伝説, シーズン, ピート
0073	10	ブルズ,NBA, ウィザーズ,MJRetirement, 引退, 復帰, シカゴ, 属す, 無断, 図表
0073	20	NBA, ブルズ, シカゴ, ウィザーズ,Basketball, 引退, 復帰, 度目, 属す, 図表
0074	5	ブナ, 条約, 林野庁, 遺産, ユネスコ, 林, 生態, バッファーゾーン, 原生, 文化
0074	10	ブナ, 遺産, 林, 原生, 林野庁, 生態, 条約, 世界, 地域, 登録
0074	20	遺産, ブナ, 原生, 条約, 世界, 生態, ユネスコ, 五箇山, 林野庁, 林
0076	5	哲学, スピノザ, ハイデガー, カント, アンチノミー, ナーゲル, カルト, ルサンチマン, 錦野, エチカ
0076	10	哲学, スピノザ, ハイデガー, カルト,ctakasi, カント, エチカ,komaba, 彼, 批判
0076	20	哲学, スピノザ, カント, カルト, ハイデガー, 論考, ゲーデル, 論理, 書物, ニーチェ

Table 2. Expansion terms (query IDs: from 0080 to 0099)

qid	R	top 10 terms
0080	5	暖色, 赤, 黄色, 青, 寒色, 逆, リラックス, 闘牛, カラー, 威圧
0080	10	暖色, 赤, 色彩, 配色, 寒色, カラー, 青, 明度, 色相, 興奮
0080	20	暖色, 興奮, 色彩, 寒色, 赤, カラー, 青, 心, 落ち着き, 赤い
0082	5	中国共産党, 開放, 改革, 体制, 人民, 転換, 国有, 国家, 重視, 全人代
0082	10	中国共産党, 改革, 開放, 政治, 政策, 転換, 体制, 人民, 天安門, 発展
0082	20	改革, 政策, 開放, 体制, 国家, 政治, 中国共産党, 外資, 成長, 移行
0084	5	aimjal, アイム,2273,27014, ラミネート, 橙,13022,13115,13086,13102
0084	10	aimjal, アイム,2273, 橙, 用命,aim,588, 組合せ, 見本,PUD
0084	20	aimjal, アイム,2273, 用命,aim,588, ラミネート, 見本, 粘着, 選定
0086	5	語学, 見つける, アメリカ, オーストラリア, 短期, ビザ, アイルランド, ホームステイ, 学校, ニュージーランド
0086	10	語学, ビザ, アメリカ, 就職, オーストラリア, ニュージーランド, 短期, ホリデー, 仕事, 滞在
0086	20	語学, オーストラリア, 就職, ビザ, ホリデー, ニュージーランド, 英語, アメリカ, インターン, ワーキング
0088	5	酔い, 取消し, 無免許, 欠格, 無事故, 違反, 公安, 処分, 仮免許, 前歴
0088	10	違反, 処分, 欠格, 停止, 事故, 公安, 取消し, 無事故, 前歴, 取消
0088	20	違反, 処分, 交通, 欠格, 累積, 停止, 事故, 前歴, 取消し, 期間
0091	5	会話,ETS, 文法,990, 英語, 能力,Chauncey, リスニング, リーディング, 語彙
0091	10	文法, 会話, 語彙,ETS, リスニング, リーディング, テスト, 受験, 英語, 力
0091	20	リスニング, 文法, リーディング, 語彙,ETS, テスト, 英語,TOEFL, 問題, 会話
0095	5	万世橋, 明治, 建造, 復元, 名主, 造り, 造営, 格式, 移築, 社寺
0095	10	復元, 明治, 建造, 移築, 年代, たて, 博物館, 造り, 至る, 万世橋
0095	20	移築, 建造, 明治, 復元, 土蔵, 造り, 年代, 大正, 茶室, 末期
0097	5	目元, メイク, まぶた, 色, アイホール, アイカラー, アイシャドウ, 目尻, まつ毛, ぼかす
0097	10	メイク, アイシャドウ, 目元, アイカラー, アイライン, 色,addd, カラ, まぶた, マス
0097	20	メイク, アイカラー, アイシャドウ, 目元, 色, アイライン, まぶた, 眉, アイシャドー, 一重
0098	5	ユネスコ, 古都, 五箇山, 建造, 厳島, 姫路城, 白川郷, 白神山地, 条約, 法隆寺
0098	10	ユネスコ, 屋久島, 条約, 白神山地, 古都, 五箇山, 厳島, 姫路城, 白川郷, 法隆寺
0098	20	ユネスコ, 条約, 白神山地, 白川郷, 屋久島, 古都, 法隆寺, 五箇山, 社寺, 厳島
0099	5	脱走, ガリ, ゴールデン, 毛並み, ジャン, あげる, ロボロフスキーハムスター, おしり, 匹, 死ぬ
0099	10	ガリ, 回し, ジャン, 匹, アン, 入れる, えさ, ペット, 飼育, あげる
0099	20	匹, 回し, あげる, ゴールデン, 飼育, ガリ, 掃除, ペット, えさ, 頬袋

Table 3. Comparision of average precisions

$k/R/T$	0.005	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.10	0.12	0.14	0.16	0.18	0.2
10/5/5	1.005	1.017	1.027	1.041	1.039	1.045	1.043	1.047	1.051	1.055	1.056	1.053	1.052	1.039	1.016
10/5/10	1.007	1.009	1.023	1.031	1.033	1.038	1.045	1.051	1.053	1.062	1.071	*1.075*	1.071	1.068	1.060
10/5/20	1.006	1.013	1.018	1.023	1.031	1.036	1.039	1.038	1.042	1.048	1.047	1.049	1.047	1.049	1.031
10/10/5	1.008	1.017	1.029	1.034	1.038	1.046	1.054	1.060	1.061	1.061	1.027	1.018	1.005	0.986	0.960
10/10/0	1.007	1.014	1.019	1.027	1.030	1.032	1.035	1.038	1.037	1.026	1.014	1.008	0.975	0.962	0.942
10/10/20	1.005	1.012	1.014	1.002	1.005	1.013	1.014	1.014	1.006	0.994	0.976	0.956	0.937	0.919	0.907
10/20/5	1.008	1.017	1.029	1.035	1.038	1.045	1.043	1.047	1.018	1.011	0.996	0.985	0.971	0.948	0.926
10/20/10	1.011	1.016	1.028	1.037	1.043	1.050	1.056	1.052	1.053	1.048	1.019	0.972	0.949	0.935	0.922
10/20/20	1.006	1.011	1.012	1.015	1.019	1.016	1.019	1.016	1.010	0.988	0.963	0.937	0.912	0.890	0.868
20/5/5	1.005	1.018	1.034	1.042	1.041	1.041	1.046	1.049	1.053	1.060	1.059	1.066	1.049	1.023	0.987
20/5/10	1.007	1.010	1.025	1.031	1.036	1.042	1.050	1.051	1.058	1.068	1.074	***1.078***	1.070	1.066	1.060
20/5/20	1.006	1.013	1.020	1.028	1.032	1.036	1.041	1.045	1.045	1.051	1.048	1.045	1.041	1.044	1.024
20/10/5	1.008	1.019	1.030	1.037	1.043	1.053	1.060	1.063	1.063	1.036	1.028	1.015	0.996	0.965	0.937
20/10/10	1.007	1.014	1.020	1.030	1.033	1.034	1.038	1.037	1.034	1.018	1.004	0.973	0.959	0.942	0.917
20/10/20	1.005	1.013	1.015	1.003	1.010	1.014	1.013	1.013	1.006	0.983	0.956	0.928	0.914	0.895	0.867
20/20/5	1.008	1.018	1.029	1.037	1.042	1.039	1.044	1.014	1.011	1.002	0.996	0.981	0.954	0.918	0.870
20/20/10	1.012	1.017	1.029	1.040	1.047	1.054	1.055	1.052	1.052	1.032	0.974	0.951	0.935	0.921	0.906
20/20/20	1.006	1.012	1.013	1.015	1.019	1.021	1.016	1.012	1.003	0.972	0.949	0.914	0.889	0.862	0.846
50/5/5	1.010	1.019	1.035	1.044	1.038	1.043	1.048	1.056	1.061	1.060	1.064	1.047	1.017	0.972	0.927
50/5/10	1.007	1.016	1.030	1.031	1.039	1.046	1.053	1.057	1.060	1.065	1.069	*1.070*	1.065	1.064	1.053
50/5/20	1.007	1.013	1.020	1.029	1.031	1.039	1.043	1.048	1.053	1.058	1.045	1.037	1.038	1.031	1.010
50/10/5	1.013	1.020	1.033	1.043	1.049	1.056	1.059	1.061	1.043	1.029	1.019	0.998	0.965	0.933	0.907
50/10/10	1.007	1.014	1.026	1.033	1.037	1.035	1.036	1.033	1.021	1.003	0.973	0.963	0.941	0.916	0.892
50/10/20	1.009	1.013	1.014	1.005	1.012	1.010	1.010	1.000	0.991	0.961	0.933	0.910	0.890	0.874	0.831
50/20/5	1.013	1.023	1.031	1.037	1.037	1.038	1.016	1.008	1.003	0.998	0.982	0.957	0.915	0.868	0.814
50/20/10	1.012	1.018	1.031	1.044	1.050	1.056	1.057	1.054	1.041	0.980	0.951	0.933	0.913	0.894	0.876
50/20/20	1.011	1.013	1.014	1.016	1.019	1.020	1.017	1.009	0.993	0.960	0.921	0.890	0.861	0.843	0.809
100/5/5	1.011	1.020	1.036	1.043	1.039	1.043	1.049	1.058	1.061	1.060	1.055	1.030	0.999	0.935	0.903
100/5/10	1.007	1.016	1.031	1.034	1.040	1.047	1.052	1.060	1.061	1.065	*1.067*	1.066	1.059	1.054	1.039
100/5/20	1.006	1.013	1.021	1.029	1.034	1.039	1.044	1.049	1.054	1.054	1.044	1.038	1.026	1.027	1.005
100/10/5	1.014	1.021	1.033	1.045	1.049	1.054	1.058	1.058	1.036	1.027	1.008	0.981	0.954	0.907	0.879
100/10/10	1.011	1.015	1.026	1.034	1.039	1.040	1.032	1.026	1.016	0.994	0.968	0.953	0.936	0.902	0.881
100/10/20	1.009	1.013	1.015	1.003	1.012	1.008	1.004	0.992	0.979	0.945	0.922	0.897	0.878	0.842	0.813
100/20/5	1.014	1.023	1.032	1.035	1.036	1.039	1.009	1.005	1.003	0.992	0.967	0.934	0.894	0.831	0.793
100/20/10	1.013	1.022	1.034	1.044	1.051	1.057	1.053	1.046	1.027	0.970	0.941	0.919	0.899	0.875	0.854
100/20/20	1.011	1.012	1.012	1.016	1.019	1.020	1.017	1.007	0.987	0.946	0.907	0.878	0.851	0.838	0.795

Table 4. Comparision of R-precisions

$k/R/T$	0.005	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.10	0.12	0.14	0.16	0.18	0.2
10/5/5	0.994	0.994	1.012	1.016	1.021	1.040	1.047	1.051	1.050	1.057	1.057	1.037	1.030	1.024	0.989
10/5/10	1.002	1.008	1.010	1.022	1.026	1.027	1.033	1.046	1.050	1.052	1.042	1.039	1.024	1.039	1.030
10/5/20	1.002	1.001	0.999	1.008	1.017	1.021	1.031	1.039	1.041	1.035	1.039	1.037	1.038	1.036	1.051
10/10/5	0.995	1.002	1.016	1.022	1.037	1.037	1.048	1.048	1.054	***1.073***	1.017	0.996	0.970	0.974	0.951
10/10/10	1.001	1.001	1.001	1.017	1.025	1.020	1.030	1.030	1.032	1.040	1.022	1.013	0.964	0.941	0.924
10/10/20	1.001	0.994	0.992	1.004	1.021	1.025	1.018	1.022	1.021	1.016	1.015	1.006	1.002	0.991	0.967
10/20/5	0.994	1.003	1.017	1.025	1.036	1.034	1.027	1.031	1.003	1.003	1.002	0.973	0.954	0.945	0.930
10/20/10	0.995	0.995	1.019	1.031	1.042	1.040	1.042	1.056	1.056	1.051	1.034	0.974	0.954	0.941	0.926
10/20/20	1.001	0.993	0.999	1.023	1.022	1.021	1.022	1.030	1.028	1.013	1.003	0.966	0.949	0.944	0.934
20/5/5	0.993	0.996	1.014	1.017	1.023	1.045	1.049	1.050	1.059	1.059	1.050	1.054	1.046	1.019	0.963
20/5/10	1.001	1.008	1.007	1.023	1.026	1.029	1.033	1.049	1.051	1.035	1.040	1.038	1.037	1.035	1.024
20/5/20	1.002	1.001	0.999	1.014	1.018	1.022	1.025	1.039	1.042	1.039	1.039	1.042	1.018	1.031	1.026
20/10/5	0.993	1.008	1.020	1.024	1.034	1.046	1.049	1.057	*1.062*	1.033	1.004	0.982	0.969	0.963	0.936
20/10/10	0.994	0.999	1.007	1.015	1.016	1.020	1.020	1.030	1.037	1.020	1.004	0.960	0.953	0.904	0.894
20/10/20	1.001	0.994	1.000	1.016	1.023	1.018	1.022	1.019	1.015	1.010	1.011	1.007	0.982	0.967	0.963
20/20/5	0.992	1.008	1.014	1.025	1.035	1.020	1.023	0.996	0.992	1.000	0.976	0.958	0.953	0.931	0.923
20/20/10	0.994	1.003	1.020	1.032	1.034	1.040	1.048	1.055	1.054	1.045	0.979	0.958	0.927	0.919	0.891
20/20/20	1.002	0.993	1.000	1.022	1.024	1.019	1.018	1.025	1.020	1.015	0.977	0.960	0.945	0.930	0.892
50/5/5	0.993	1.005	1.015	1.018	1.036	1.046	1.046	1.053	1.063	1.058	*1.064*	1.050	0.992	0.964	0.931
50/5/10	1.001	1.008	1.008	1.017	1.025	1.028	1.029	1.028	1.048	1.046	1.041	1.048	1.038	1.026	1.015
50/5/20	1.002	1.001	1.004	1.014	1.018	1.023	1.024	1.043	1.040	1.038	1.035	1.027	1.028	1.019	1.010
50/10/5	1.000	1.008	1.018	1.032	1.040	1.044	1.040	1.050	1.047	1.002	0.978	0.977	0.965	0.948	0.904
50/10/10	0.994	1.000	1.009	1.010	1.017	1.019	1.024	1.025	1.026	1.009	0.982	0.935	0.896	0.875	0.852
50/10/20	1.002	0.994	1.004	1.015	1.011	1.015	1.010	1.013	1.013	1.002	0.997	0.992	0.965	0.945	0.905
50/20/5	0.994	1.008	1.016	1.029	1.034	1.020	1.022	0.999	0.998	0.968	0.960	0.961	0.936	0.909	0.870
50/20/10	0.995	1.003	1.021	1.025	1.025	1.043	1.051	1.056	1.039	1.024	0.979	0.928	0.934	0.897	0.876
50/20/20	1.001	0.992	1.000	1.023	1.017	1.016	1.021	1.032	1.018	1.003	0.957	0.956	0.922	0.896	0.849
100/5/5	0.993	1.006	1.015	1.026	1.038	1.042	1.041	1.060	*1.062*	1.056	1.057	1.016	0.965	0.936	0.893
100/5/10	1.001	1.008	1.008	1.019	1.026	1.024	1.029	1.039	1.045	1.047	1.060	1.044	1.033	1.025	1.024
100/5/20	1.002	1.001	1.004	1.016	1.020	1.023	1.026	1.043	1.036	1.038	1.023	1.032	1.018	1.010	1.000
100/10/5	1.000	1.009	1.018	1.038	1.041	1.041	1.044	1.049	1.053	0.997	0.970	0.968	0.948	0.909	0.868
100/10/10	0.994	1.000	1.009	1.011	1.021	1.020	1.024	1.024	1.007	0.987	0.941	0.932	0.881	0.860	0.833
100/10/20	1.002	0.994	1.004	1.015	1.010	1.015	1.009	1.012	1.019	1.001	0.994	0.971	0.947	0.927	0.889
100/20/5	1.001	1.009	1.016	1.029	1.028	1.027	1.001	0.997	1.003	0.970	0.956	0.942	0.935	0.883	0.825
100/20/10	0.995	1.005	1.021	1.022	1.034	1.039	1.050	1.044	1.034	0.987	0.966	0.921	0.896	0.885	0.871
100/20/20	1.001	0.992	1.012	1.023	1.017	1.014	1.017	1.021	1.021	0.993	0.949	0.943	0.908	0.858	0.839

Table 5. Web search improvements characterized by the correlations between query term RSVs and expansion term RSVs (See also Figure 1.)

query ID	0001	0003	0004	0006	0019
(average precision after QE) − (baseline average precision)	0.0402	0.0196	0.0002	0.0043	-0.0066
MAX(expansion term RSVs) − MAX(query term RSVs)	-0.1558	-2.3770	-0.9446	3.0651	-3.1196
STDEV(expansion term RSVs) − STDEV(query term RSVs)	0.5141	-3.3720	-0.2685	-0.6734	-0.5968

0021	0022	0023	0028	0029	0034	0044	0045	0055	0058	0061
0.1958	0.0235	0.0000	0.0064	0.0106	0.0263	-0.0176	0.0435	-0.0097	-0.0042	-0.0101
0.4644	-3.1396	1.0490	-5.0153	0.1731	-0.6875	-0.0916	-2.6462	-2.4513	-2.1042	-0.4227
0.4119	-3.6928	-0.0317	-1.1549	-0.4196	-1.5490	-0.0990	-0.8502	-1.9339	-0.4612	-0.1777

0062	0063	0065	0068	0070	0071	0073	0074	0076	0080	0082
0.0107	0.0059	-0.0014	0.0033	0.0199	-0.0020	-0.0004	-0.0154	-0.0056	0.0279	0.0146
0.4745	0.1828	-1.8797	-2.3629	0.2741	3.0469	-1.1962	-1.5094	-5.3382	-0.3633	1.9622
-1.0617	-0.7537	-1.1954	-1.4337	1.1591	-0.8196	0.4171	-1.4308	-4.4108	-0.7170	0.3942

0084	0086	0088	0091	0095	0097	0098	0099
-0.0002	-0.0013	0.0155	-0.0060	0.0049	-0.0788	0.1134	0.0128
3.3783	0.3146	1.3507	-2.1818	-1.9612	-1.3420	1.8927	-1.7562
-0.5581	-0.0492	-0.1243	-2.9149	-0.7062	-5.1956	-0.9571	-2.0401