1E2-3

キーワードと日時を入力とした検索における関連語の獲得

Related Term Extraction in Temporal Information Retrieval

廣嶋 伸章[*1] 別所 克人[*1] 小池 義昌[*1] 片岡 良治[*1]
Nobuaki Hiroshima Katsuji Bessho Yoshimasa Koike Ryoji Kataoka

[*1]日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

We have been developing an information retrieval system that accepts temporal queries in addition to ordinary keywords. Such a system is useful, for example, when a user want to find information about events held at a certain time. However, the user sometimes have difficulty grasping the outline of the vast amount of results. Showing representative terms related to the query is an effective way to solve the problem. We propose a method that selects related terms based on their co-occurrence relation with the termporal expression that matches the specified time as well as their relavancy to the query keywords.

1. はじめに

情報検索は行動の指針を決定する際などに幅広く利用され、人々の生活に必要不可欠なものとなっている。人々の検索のニーズは多様であるが、その中には通常のキーワードに加えて日時を指定したいという場合が少なからず考えられる。例えば、クリスマスイブに行われたイベントについて知りたい場合には、キーワードとして「イベント」、日時として「2010 年 12 月 24 日」を指定して検索を行いたいような場合が想定される。しかしながら、通常のキーワードによる検索と同様に、キーワードと日時を指定した検索においても、検索結果が膨大となった場合にどのような情報が検索結果の中に含まれているかの概要を把握することが困難という問題が生じる。この問題を解決するための方法の一つとして、検索結果の文書に含まれキーワードに関連を持つ語 (以下では関連語と呼ぶ) を提示するという方法が挙げられる。特に、キーワードと日時を入力とした検索においては、入力の日時に特有の関連語を提示することが効果的であると考えられる。例えば先ほどのイベントに関する検索では、12 月 24 日に行われる実際のイベント名などが関連語として考えられる。また、キーワードが「映画」の場合は、その日に上映開始の映画タイトルや、その日に行われる試写会に登場する俳優名などが関連語として考えられる。このような関連語が提示されれば、指定した日時に関する検索結果の概要を把握することが可能となるだけでなく、関連語を選択して検索結果の絞り込みを行うことが可能となる。そこで、本研究では、キーワードと日時を入力とした検索において検索結果の文書に含まれる関連語を獲得することを目的とする。

2. 関連研究

検索結果からの関連語の獲得に関する研究として、検索結果の絞り込みのための語を提示する研究や、検索結果のクラスタリングにおいてクラスタの内容を表す語を提示する研究が行われている。酒井らは、TF-IDF に加えて語が多くの文書に分散して出現しているかどうかの指標を用いて絞り込み語を獲得する手法を提案している [酒井 00]。Zeng らも、TF-IDF やフレーズの独立性に関する指標を用いて回帰分析によりフレーズのランキングを行う手法を提案している [Zeng 04]。

連絡先: hiroshima.nobuaki@lab.ntt.co.jp

キーワードと日時を入力とした検索における関連語の獲得では、関連語が指定した日時に関連するかどうかが重要視される。従来手法では、例えば「映画」というキーワードに対する検索結果の文書に「映画館でポップコーンを食べた」というような記述がある程度含まれていると、「ポップコーン」が関連語として獲得されてしまうが、このような日時に関連のない語は関連語として不適切であると考えられる。

3. 提案手法

提案手法では、キーワードと日時の関連性を考慮したランキングおよびスニペット生成を行い、スニペットから関連語候補の抽出を行う。得られた関連語候補のそれぞれが入力された日時に関連しているかどうかの判定を行い、日時に関連している語を関連語として提示する。

3.1 ランキング

キーワードと日時の関連性を考慮したランキングでは、過去に提案した有効範囲を考慮した手法 [廣嶋 10] を用いる。

3.2 スニペットの生成

スニペットの生成では、検索結果が上位の文書を文に分割し、上記のランキングにおけるスコアに大きく貢献したキーワードおよび日時を含む文を選択してスニペットとする。

3.3 関連語候補の抽出

関連語候補の算出では固有表現に限らず任意の名詞句を関連語候補として、スニペットから抽出する。任意の名詞句とした理由は、イベント名のようにうまく固有表現として抽出できない語句や、一般名詞で構成される語句も関連語となる可能性があることを考慮したためである。ここでは、Suffix Array の考え方に基づき、以下の手順で関連語候補を抽出する。

1. 各単語から始まる文の生成
2. 生成された文のソート
3. 名詞句の頻度の算出
4. 頻度による名詞句の選択

各単語から始まる文を生成し、それらの文を辞書順にソートすることにより、同一の名詞句から始まる文が連続するようになる。そのため、各文が前の文と文頭から何単語一致するか

を調べておき、長さ N のある名詞句から始まる先頭の文から N 単語以上一致する文がいくつ連続するかをカウントすることにより、その名詞句の頻度を算出することができる。

頻度による名詞句の選択では、頻度に加えて名詞句の長さも利用する。これは、長い名詞句ほど出現しにくいことを考慮したものである。以下の式により閾値 T_w を求め、頻度が閾値以上の名詞句を関連語候補として抽出する。

$$T_w = \frac{\alpha|D|}{l_w} + \beta$$

ここで、$|D|$ は検索結果の上位の文書の数、l_w は名詞句 w の単語数、α, β は定数である。

3.4 日時関連性の判定

日時関連性の判定では、関連語候補の出現と入力された日時の出現との間に関連があるかどうかを表すカイ 2 乗値を求め、値が高いものを関連語として獲得する。カイ 2 乗値は以下の式により求めることができる。

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

ここで、a は関連語候補を含み入力の日時を含む文書の数、b は関連語候補を含み入力の日時を含まない文書の数、c は関連語候補を含まず入力の日時を含む文書の数、d は関連語候補を含まず入力の日時を含まない文書の数である。これらの文書数は、検索エンジンを用いて検索を行うことにより取得する。

4. 評価

提案手法の有効性を検証するため、評価を行った。

4.1 実験条件

検索対象の文書として、2011 年 2 月までの 1 年間に投稿されたブログ記事を利用した。各記事に対して抽出された日時表現を日時に変換して検索インデックスを作成した。入力のキーワードとして「イベント」「映画」の 2 種類を用意し、各キーワードに対して 3 種類ずつの日時を用意して合計 6 回の検索を行った。検索結果の上位 1000 件を取得し、$\alpha = 0.005, \beta = 2$ として提案手法により 30 ずつの関連語を獲得した。

4.2 従来手法との比較

まず、従来手法との比較を行った。従来手法として、[酒井 00] の手法を用意した。それぞれの手法で獲得された関連語のそれぞれについて、1 人の評価者により以下の観点で評価を行った。

- 妥当性 · · · 関連語として妥当かどうか
- 有用性 · · · 関連語が検索結果の絞り込みに有効かどうか
- 日時関連性 · · · 関連語が日時に関連しているかどうか

それぞれの観点について 1 または 0 の点数を付与し、上位 10 語、20 語、30 語での平均を求めた。有用性については、提示された関連語で絞り込みを行い、検索結果が 4 件以上 500 件以下となった場合に 1 とした。結果を表 1 に示す。

結果より、いずれの観点においても従来手法に比べて正しく関連語が獲得できることが確認された。

表 1: 従来手法との比較評価結果

観点	手法	10 語	20 語	30 語
妥当性	従来手法	0.333	0.308	0.328
	提案手法	0.817	0.733	0.689
有用性	従来手法	0.350	0.300	0.278
	提案手法	0.783	0.667	0.622
日時関連性	従来手法	0.333	0.275	0.261
	提案手法	0.767	0.667	0.595

4.3 考察

獲得された関連語について考察を行う。

キーワードとして「イベント」を入力とした場合には、イベントの種類を表す語 (「チャリティーイベント」など) やイベントの内容を表す語 (「THIS IS IT」など) が多く獲得された。また、キーワードとして「映画」を入力とした場合には、入力した日時に行われている映画祭の名称 (「東京国際映画祭」など) やその日から上映が始まる映画のタイトル・出演俳優の名前が多く獲得された。このことから、関連語の獲得は適切に行われていると考えられる。

妥当性の観点で不適切と判定された例としては、「イベント開催」「全国公開」のように、日時によらずキーワードと密接に関係する語が存在した。このような語に対しては、語を含む文書で述べられている日時の分布を調べ、分布が一様であるものについては除去する必要がある可能性があると考えられる。

有用性の観点で不適切と判定された語は、特定の文書での出現頻度が高いが、多くの文書に出現しないため、絞り込みに有効でないと考えられる。頻度をカウントする際に同一文書内で出現した語は重複してカウントしないようにすることで、除去できる可能性があると考えられる。

日時関連性で不適切と判定された語は、数多くの日時について述べられている文書の中でよく出現していることが確認された。数多くの日時について述べられている文書は処理の対象としないようにすることで、日時に関連しない語が除去できる可能性があると考えられる。

5. まとめ

キーワードと日時を入力とした検索において、検索結果のスニペットから任意の名詞句を関連語候補として抽出し、入力の日時との関連性を判定することにより関連語を獲得する手法を提案した。提案手法の有効性を検証するための評価を行った結果、妥当性・有用性・日時関連性のいずれにおいても従来手法に比べて正しく関連語が獲得できることが確認された。

今後は、考察で述べた方法により、それぞれの観点に関して精度を向上させていきたいと考えている。

参考文献

[酒井 00] 酒井浩之, 大竹清敬, 増山繁: 絞り込み語の提示による検索支援の試み, 言語処理学会第 6 回年次大会 (2000).

[Zeng 04] Zeng, H .J., He, Q .C., Chen, Z., Ma, W. Y. and Ma, J.: Learning to cluster web search results, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04), pp.210-217 (2006).

[廣嶋 10] 廣嶋伸章, 別所克人, 小池義昌, 片岡良治: 記述された日時の有効範囲を考慮した日時指定検索, WebDB Forum 2010 (2010).