Web検索のための有用な関連キーワードを評価するシステムの構築

寺島浩太† 安藤大幸† 藤本敬介† 中山泰一†
†電気通信大学 情報工学科

1 はじめに

Web 検索において検索結果のページが大量に発見された場合，目的の情報を探し出すには労力を要する場合がある．この問題を解決するために，ユーザに関連キーワードを提示する検索支援システムがある．しかし，既存のシステムではユーザの検索履歴を参考にしているため，誤植などの絞込みに適していない関連キーワードが提示される可能性があり，どのキーワードが最も有用であるかがわからない．そこで，本研究ではユーザの要求に応じた最も有用な関連キーワードを評価するシステムの構築を行った．結果数を減らし，有用な情報を得たいというユーザの要求に対し，本研究では絞り込み検索に用いる関連キーワードの有用度を定義することにより，関連キーワードが情報を適切に限定できるかどうかの相対的な評価を行う．実験により本手法が良好に機能することを示す．

2 関連サービス, 関連研究

ユーザに関連キーワードを提示する検索支援システムの代表として，Google サジェスト [1] がある．ユーザに関連キーワードを提示する検索支援システムの研究として，野田 [2] らが提案した，主題語から話題語を抽出し，提示するシステムがある．

3 キーワードの評価

3.1 評価の考え方

キーワード a を b で絞り込みたい時，有用な絞込みの条件は以下のようになる．

1. b は a に似ていない
2. b にとって a は主要な情報である

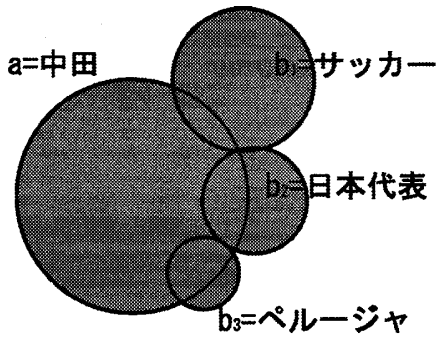

図 1: 検索目的:中田英寿選手, a = 中田, b_1 = サッカー, b_2 = 日本代表, b_3 = ペルージャ (簡単のため $a \cap b_1 \cap b_2$ などの積集合は考えない)

1 について，類義語，同義語の類や，検索結果が似通っているキーワードをを AND 検索しても，検索結果を絞り込むことはできない．2 について，ベン図で表すと図 1 のようになる．図 1 のケースでは b_1 から b_3 へ向かうほど中田英寿選手選手という個人を特定しやすい．つまり，b の検索結果の集合が小さくて，b にとって a が主要な情報であるほうが有用な絞込みであると考えられる．

この考え方における「似ている」という概念を数値化したものを類似度，b にとって a が主要な情報であるという概念を数値化したものを b の a に対する依存度として定義する．

3.2 評価値の計算

キーワード a を b で絞り込みたい時，a の検索結果の Web ページ群と b の検索結果の Web ページ群の中に同一のページが存在するのならば，類似度が高いと考えることができる．また，その数が多ければ多いほど類似度がより高いと考えることができる．

b の検索結果の件数のうちの a AND b の検索結果の件数の割合が高ければ，b の a に対する依存度が高いと考えることができる．また，逆も同様である．

Evaluation method of related keyword
Kouta Terashima†, Hiroyuki Ando†, Keisuke Fujimoto†and Yasuichi Nakayama†
†Department of Computer Science, The University of Electro-Communications.

これらを踏まえて，キーワード a を b で絞り込みたい時，キーワード x の検索結果の Web ページ群を $f(x)$，キーワード x の検索結果の件数を $g(x)$ とし，類似度 sam,a の b に対する依存度 dep_a,b の a に対する依存度 dep_b を以下のように定義する．

$$
\begin{aligned}
sam(a,b) &= |\{x|x \in f(a), y \in f(b), x = y\}| \\
dep_a(a,b) &= g(a \text{ AND } b)/g(a) \\
dep_b(a,b) &= g(a \text{ AND } b)/g(b)
\end{aligned}
$$

この定義に基づき計算を行い，類似度が高すぎず，b の a に対する依存度が高いものを有用と評価する．

4 システムの設計

4.1 情報の適切な限定，拡張

基準となるキーワードの組み合わせを a_s, b_s とし，その後入力したキーワードを a_i, b_i とする．類似度の基準値を R とし，$sam(a_i, b_i) > R$ となった場合は有用でない絞込み検索であるとする．$sam(a_i, b_i) \leq R$ の場合について，情報の適切な限定，拡張，情報の限定，拡張を以下のように定義する．

適切な限定	限定
$dep_a(a_i,b_i) < dep_a(a_s,a_s)$ $dep_b(a_i,b_i) \geq dep_b(a_s,a_s)$	$dep_a(a_i,b_i) < dep_a(a_s,a_s)$ $dep_b(a_i,b_i) < dep_b(a_s,a_s)$
適切な拡張	拡張
$dep_a(a_i,b_i) > dep_a(a_s,a_s)$ $dep_b(a_i,b_i) \geq dep_b(a_s,a_s)$	$dep_a(a_i,b_i) > dep_a(a_s,a_s)$ $dep_b(a_i,b_i) < dep_b(a_s,a_s)$

4.2 システムの概要

実験，評価が行いやすいように，有用な関連キーワードを評価し，提示するシステムを実装した．ユーザが入力したキーワードの組み合わせ a, b(2 語の AND 検索を想定) に対し，Yahoo!関連検索ワード [3] を dep_b の値で降順にソートして，dep_a の値により，情報の適切な限定，情報の限定，情報の適切な拡張，情報の拡張の 4 つに分けて提示する．このシステムにより，ユーザは Yahoo!関連検索ワードの中から目的とする情報量に合った有用なキーワードを選択することができる．

4.3 システムの動作

1. ユーザがキーワード "$a\ b_i$" を入力．
2. $sam(a, b_i)$ を計算し，$sam(a, b_i) > R$ ならばユーザに提示．
3. $dep_a(a, b_i), dep_b(a, b_i)$ を計算した後，ファイルから読み込んだ $dep_a(a, b_1)..dep_a(a, b_n)$, $dep_b(a, b_1)..dep_b(a, b_n)$ と比較していき，関連検索ワードを 4 つに分けて提示する．

表 1: a = 中田, b = 日本代表 とした時提示される関連検索ワード
(簡単のため, 明らかに無関係と思われる関連検索ワードは除く)

適切な限定	限定	適切な拡張	拡張
中田 ペルージャ 中田 ヒデ	中田 浩司 中田 英俊 中田 英 中田 ひで	なし	中田 サッカー 中田 氏 中田 ブログ

5 実験と評価

キーワード a AND b を入力し，提示される関連キーワードを確認する．これをいくつかの a, b の組み合わせに対して行った．

a = 中田, b = 日本代表 とした時に提示される関連検索ワードを表 2 に示す．適切な限定の欄には "ペルージャ" や "ヒデ" など，中田英寿選手に特定できるキーワードが提示されている．限定の欄には "浩司" や "英俊" などの誤植や，"ひで" や "英" など，"ヒデ" よりも中田英寿選手と関連性が低いと思われるキーワードが並んでいる．適切な拡張については，該当するキーワードはなかった．これは，b よりも多くの検索結果が得られて，有用な絞込みになるキーワードが関連検索ワードの中になかったことを示している．拡張の欄には中田浩二選手や横浜市長の中田宏氏などに情報が分散してしまう "サッカー"，"氏"，"ブログ" などのキーワードが並んでいる．これらの結果により，概ねユーザの要求する情報量に一致する有用なキーワードが提示されることがわかった．

6 まとめ

本研究では，ユーザの要求する情報量に応じた最も有用な関連キーワードを評価するシステムの構築，実験による評価を行った．今後の課題としては，3 語以上のキーワードの組み合わせに対する対応，人物や事物を特定することが目的でない AND 検索への対応などが考えられる．

参考文献

[1] Google サジェスト
http://www.google.co.jp/webhp?complete=1&hl=ja

[2] 野田武史, 大島裕明, 小山 聡, 田島敬史, 田中克己: 主題語からの話題語自動抽出とこれに基づく Web 情報検索, 日本データベース学会 Letters Vol.5, No.2, pp.69-72(2006).

[3] Yahoo!関連検索ワード Web サービス
http://developer.yahoo.co.jp/search/webunit/V1/webunitSearch.html