テキストデータベースのための文書ランキング法

3S−2

小川 泰嗣 別所 礼子 西村 美苗 広瀬雅子
(株) リコー 研究開発本部 中央研究所

1 はじめに

われわれはキーワードに基づくテキストデータベース管理システムを研究開発中である [1][3]。キーワードに基づくシステムでは、検索要求はユーザが入力した検索語と登録テキストの索引語の一致から検索結果が生成される。そのため、検索語と索引語が部分一致しても検索できないシステムが多い。同義語辞書によって部分一致検索を実現することも可能であるが、日本語には複合語が多いため全てを辞書に登録することは非現実的である。

これに対し、テキスト登録時に索引を索引語の構成単語に基づいて作成し、検索時には検索語も構成単語に分解して検索処理を行なえば部分一致検索を簡単に実現できる [2]。しかし、この方式にはテキスト登録時のランキング法に改善の余地があった。そこで、ランキングルールを改良するとともに、キーワード抽出用のキーワード素性を用いることで高精度な文書ランキングを実現した。

2 文書ランキング法

2.1 概要

ユーザは検索対象を検索語として入力する。システムは、テキストごとにそのテキストに付与されている索引語から得点を計算し、ランキングを行なう。

1. キーワード抽出：
   [3] の方法にしたがい入力検索語から実際に検索に用いられる単語が選択される。ただし、抽出ルールは一部異なっている。
2. 重要度付与：
   重要度とは、検索語の形態素解析した結果得られる各単語に付与される各単語の重要性を表す値。後述するルールに従って各単語ごとに重要度を計算する。
3. 一致度計算：
   一致度とは、登録文書に付与されている各索引語と検索語の一致の程度を表す値。検索語の各単語の重要度から計算される。
4. 文書得点計算：
   文書得点とは、登録文書と検索語の一致の程度を表す値。登録文書に付与されている各索引語と検索語の一致度から計算される。

以下では、2 〜 4 の処理を詳細に説明する。

A Text Ranking Method for Text Database Systems
Yasushi Ogawa, Ayako Bessho, Mina Nishimura and Masako Hirose (Research & Development Center, RICOH Co., Ltd.)

2.2 重要度付与

検索語は複数の単語から構成される複合語であることが多い。複合語は文書ランキングにおけるキーワード抽出で単語に分割されるため、部分一致検索は可能となる。しかし、構成単語の重要性は複合語の品詞・位置などによって異なる。そこで、検索語の構成単語ごとにつぎのルールにしたがって重要度を付与する。

1. 検索語において最も語尾に近い品詞群 1 の単語の重要度は基本点とする。
2. それ以外の品詞群 1 の単語の重要度は、その位置より最も近い後方にある品詞群 1 の重要度に増加分を加えた値とする。
3. 「接頭修飾」付きの接頭辞の重要度は基本点とする。
4. 「接頭修飾」なしの接頭辞の重要度は 0 とする。
5. 品詞群 2 の単語の重要度は、(1) 品詞群 1 の重要度の合計 (2) 接頭修飾付の接頭語の重要度 (3) その位置より後方にある品詞群 2 の重要度の合計の 3 つを合計した値とする。
6. 上述以外の単語の重要度は 0 とする。

ここで、接頭修飾はキーワード素性の 1 つであり、キーワード抽出に用いたものと同じである [3]。また、品詞群 1 とはつぎの品詞である。

- キーワード素性付きの名詞類
- 数詞
- 接尾辞

品詞群 2 とはつぎの品詞である。

- キーワード素性なしの名詞類
- 未登録語

2.3 一致度計算

検索語の各単語の重要度をもとに検索語と検索語の一致度はつぎのように計算される。

1. 索引語に含まれる単語と一致する検索語の単語の重要度の積を一致度とする。
2. 索引語に含まれる単語並びと検索語に含まれる単語並びが一致ごとに一致度に隣接点をかける。
3. 索引語と検索語が完全一致する場合に一致度が一定となるように正規化する。

2. により、索引語に含まれる単語並びと検索語に含まれる単語並びが一致する場合に一致度が大きなり、語順の異なりを区別できる。3. により、索引語と検索語が完全に一致する際の一致度が検索語の構成単語数に依存しなくなる。

表 1: 検索語への重要度付与

	品詞	キーワード素性	重要度
新	接頭辞	接頭修飾	2
素材	一般名詞		8
研究	サ変名詞	複合語語基	3
開発	サ変名詞	複合語語基	2

2.4 文書得点計算

登録文書に付与されているキーワード数が多くても文書得点が高くなることがないようにした。テキストの各キーワードと検索語の一致度の最大値を文書得点とする。

3 例

検索語として「新素材研究開発」とする。検索語の構成単語の重要度は表 1 のようになり、正規化用の検索語点は $(2\times8\times3\times2)\times(2\times2\times2)=768$ となる。

索引語として「半導体レーザ開発」「新素材研究」が付与されているテキストを想定する。まず、索引語と検索語の一致度を計算する。「半導体レーザ」では、「開発」のみが一致するのでもと点が 2、正規化して $2\times(1000/768)=3$ 点となる。「新素材研究」では、単語として「新」「素材」「研究」が一致、単語並びとして「新／素材」「素材／研究」が一致している。したがって、もと点が $(2\times8\times3)\times(2\times2)=192$、正規化して $192\times(1000/768)=249$ 点となる。このテキストの文書得点は、「新素材研究」の一致度である 249 点となる。

4 評価

4.1 検索精度の評価

評価対象は新聞記事 200 件。評価に用いた検索語は表 2 のものである。評価基準には、再現率（検索洩れの少なさ）と適合率（ノイズの少なさ）を用いた。その際、全検索結果および上位 10 位のみから評価値を計算した。また、検索結果の判定は人手で行なった。

評価結果は表 2 に示す。再現率と比べ適合率がやや低く、重要度付与ルール、一致度計算法等の改良が望まれる。

表 2: 検索精度の評価結果

検索語	全検索結果		上位 10 位	
	再現率	適合率	再現率	適合率
ＢＳ	100	100	100	100
景気	100	100	100	100
ＮＴＴ	100	100	100	100
パソコン通信	100	22	67	40
品質管理	100	45	100	30
バブル経済	100	32	86	60
NKK 半導体開発	100	28	45	90
抗ガン剤	100	82	100	90
29 型大テレビ	100	50	100	100
新素材研開	100	29	57	40
平均	100	60	86	75

表 3: 検索速度の評価結果

絞り込み数	検索速度 (text/sec)	
	cold start	hot start
1000	1180	5000
2000	1000	2500
4000	830	1250
6000	690	833
8000	588	625

4.2 検索速度の評価

評価対象は新聞記事 20000 件、検索語は 200 語。使用マシンは SUN SPARCstation2（メモリ 32MB、内蔵 SCSI ディスク）である。検索速度はプリサーチを含めた時間から計算した。また、ディスク上のデータがキャッシュされていない状態（cold start）とキャッシュされている状態（hot start）で測定した。

測定の結果、検索時間はプリサーチによって絞り込まれたテキスト数に依存していることがわかった。それより絞り込み数に対する検索速度を計算したものが表 3 である。絞り込み数が増えるにともない検索速度が遅くなっている。しかし、20000 件のテキストならば絞り込み率 20 %（絞り込数 4000）としても、 cold start で $20000/830=24$ 秒、hot start なら $20000/1250=16$ 秒で検索でき、十分高速である。

5 おわりに

従来のキーワードに基づくテキストデータベース管理システムでは、検索語／索引語が複合語の場合に必要なテキストが検索できないという問題があった。そこで、この問題点を改善すべく、つぎのような改良を行なった。複合語を単語単位で処理することで、部分一致検索を実現し、検索の再現率を向上できた。キーワード素性を用いたランキングによって、検索時の適合率を向上できた。また、検索速度もプリサーチを導入することで、十分高速にできた。

参考文献

[1] 岩崎雅二郎他. テキストデータベースのための文字成分表を用いたプリサーチ. 第 45 回情報処理学会全国大会予稿集, 1992.

[2] 今郷詔他. 短単位キーワードを用いた文書ファイリングシステム. 第 43 回情報処理学会全国大会予稿集, 1991.

[3] 別所礼子他. テキストデータベースのためのキーワード抽出法. 第 45 回情報処理学会全国大会予稿集, 1992.