意味検索結果からのキーワードによる絞り込み効果の評価

大倉 清司　　潮田 明

(株) 富士通研究所

1. はじめに

一般的な情報検索において、ユーザーの意図通りに検索結果が出力されることは多くない。その場合、ユーザーは出力された検索結果を効率よくブラウズするために絞り込みを行うことが多い。しかし、検索結果が非常に多い場合、絞り込みを行ったとしても、目的とする文書にたどりつくまでにやはり多くの検索結果に目を通さなければならない。

自然言語処理における意味処理の効果を検証するため、我々は意味に基づくテキスト検索システムを試作した[1]。これは、検索対象文書の意味構造をデータベースに格納しておき、クエリーの意味構造に近いと思われる文書をランキング検索するシステムである。意味処理を使えば、(例えば形態素解析処理のみを使った)意味処理を使わない検索システムより高い検索精度が期待できる。例えば意味解析を使用すると、「単語を修正する」というクエリーで検索したとき、「単語」が「修正する」の対象となっている文を含む文書をマッチさせることができる。「単語」「修正」を同時に含む文書よりもユーザーの意図を反映した文書を検索できる。

今回、情報検索における意味処理の有効性を検証するのに、意味検索システムの検索結果からのキーワードによる絞り込みの効果に着目した。条件をそろえるために、意味処理を使わない、従来のキーワードベースの自然文入力の検索システムと比較した。その結果、従来のキーワードベースの自然文入力検索の検索結果からキーワードによる絞り込みを行う場合に比べ、意味検索の検索結果から絞り込むほうが絞り込み効果が高いことがわかった。

2. キーワードベース検索システム

本稿でのキーワードベース検索システムとは、キーワードを検索キーとする一般的なキーワード検索システムではなく、自然文を入力として検索する、形態素をベースとした検索システムである。文書を形態素のベクトルで表現し、クエリーから生成されたベクトルとの近似値を評価値として文書をランキングする手法は、自然文を入力して検索する一般的な手法として知られている[2]。つまり、形態素 m_1 … m_n があるとき、文書 Di のベクトルは以下のようにして表現される：

Di={Val(m_1,i),Val(m_2,i), …, Val(m_n,i)}

ただし、Val(m_k,i) = tf(m_k,i)・idf(m_k)
tf(m_k,i)は m_k の Di 内の出現回数、
idf(m_k)はデータベース内の m_k の逆文書頻度

クエリーも同様にベクトルで表現でき、データベース内の文書ベクトルとのコサイン尺度を近似値とすることができる。近似値が高いほど、クエリーに対する類似度が高いと言える。検索に不要な形態素（記号、助詞、逆文書頻度が低い形態素など）をストップワードとして除外してベクトルをつくると検索精度がよくなることが知られている。本稿のキーワードベース検索システムも、ストップワードを使い、上の手法で構築した。

3. 意味検索システム

文書が表す意味を、文書中に含まれる文の意味構造の総和と定義することができる。本研究における意味構造とは格文法[3]に基づいたグラフ構造により表現される。すなわち、意味構造は、単語の概念を表すノードおよびノード間の関係を表すアークからなる有向グラフにより表される。図1は、「太郎は花子に本をあげた。」の意味構造を有向グラフにより表した1例である。

図1において、○で囲まれたものがノードを表し、□で囲まれたものがアークを表す。ノードは、”GIVE”,”HANAKO”,”TARO”,”BOOK”の4つである。アークは”中心”,”目的”,”動作主”,”対

象”,”述語”,”過去”の 6 つである。アークはノード間の関係を表す。1 ノードにしかつながらないアークはそのノードの属性を表す。

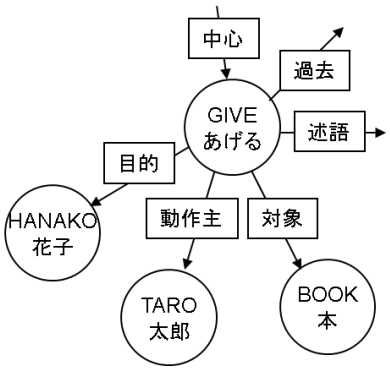

図 1. 意味構造

意味を有向グラフとして表現したとき、意味検索はグラフマッチングとして考えることができるが、単純なグラフマッチングとした場合、グラフの複雑性、意味解析処理の精度の問題がある。そこで、本稿で使用するシステムは、意味構造を、アークとそのアークにつながるノード(1 つか 2 つ)の部分グラフに分解し、この部分グラフにより検索するアルゴリズムを実装した[1]。意味構造の抽出には、日英翻訳エンジン ATLAS[4]の翻訳過程から意味構造を取り出すことにした。ATLAS は中間言語方式の翻訳方式を採用しており、原文を辞書と文法規則に基づき解析して意味構造を計算する。

4. 評価実験

2 つのシステムでの絞り込みの効果を比較するため、検索の課題とその正解をあらかじめ用意しておき、2 つのシステムに同じ自然文を入力してランキング検索した後に、同じ式で絞り込みを行うことにした。インターネット検索のような汎用的な用途の検索で評価するためには正解文書を用意することが難しく、またデータベース構築のために膨大な文書が必要であるため、今回の評価実験には向かない。著作権上の取り扱いの容易さおよびテキストデータと正解データの入手の容易性から本研究では特許を対象に 2 つのシステムを評価することにした。

実験のための検索対象データとしては、公開されている特許明細書（限定された分野のもの、約 30 万件）を使用した。課題として与えられるのは、著者らの所属機関内の公知例調査にかける前の特許アイデアを説明した書類（特許アイデア書類と呼ぶ）である。特許アイデア書類をもとに、公知例調査を行った結果提示された案件（複数あることもある）を正解とする。被験者は、特許アイデア書類を読み自然文のクエリーを作成する。その後、正解文書に含まれると思われる絞り込み用キーワードを考えておく。その後、2 つのシステムでそれぞれ、ランキング検索したときの正解文書順位とランキング検索＋絞り込み検索をしたときの正解文書順位を出す。なおクエリーは 1 つとは限らず、複数つくってもよいものとする。その場合、各クエリーに対する絞り込みキーワードも同時に考える。この流れを図 2 に示す。

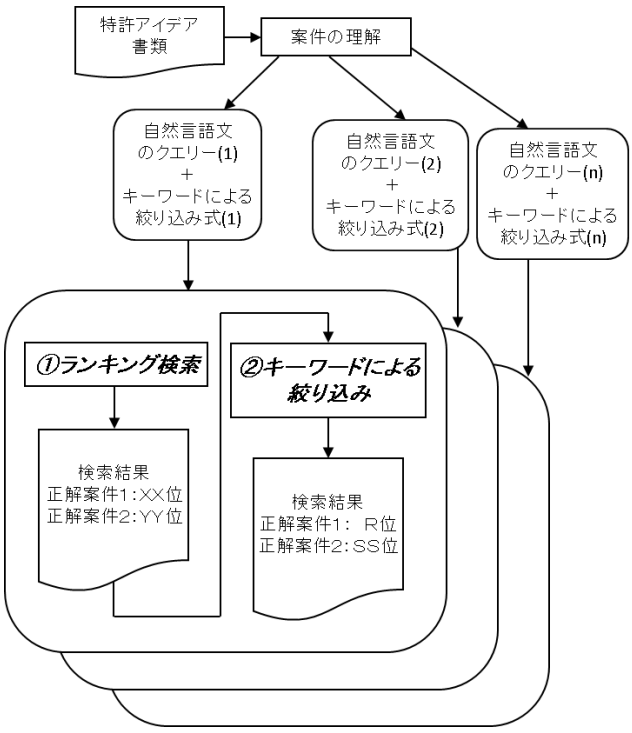

図 2. 評価の流れ

ランキング検索、絞り込み検索で各正解案件に対する順位が出るが、ランキング検索結果の正解文書順位をR、ランキング検索＋絞り込み検索の結果の正解文書順位をＢとしたときの、絞り込みの効果を以下の ***S*** を指標として評価した。

$$S=\log(R/B)$$

5. 絞り込み効果

今回の実験で得られた結果を図 3 に示す。6 つの課題について、1～2 つのクエリーを作成し、正解文書の順位を出した。絞り込み効果の平均値は意味検索が 0.837、キーワードベース検索が 0.457 で、意味検索のほうが絞り込み効果が高いことがわかった。意味検索、キーワードベース検索のいずれかの正解文書順位が 100 位以内の場合、絞り込みの効果は最大でも 2 であるため、これを除いた平均値も算出した。その場合でも、意味検索が 0.806 なのに対し、キーワードベース検索では 0.416 であった。これにより、意味検索のほうがキーワードベース検索に比べて絞り込み効果が高いことがわかった。

斜体で示される順位は、ランキング検索または絞り込み検索において、他方より正解文書順位が高いことを示す。2 つの正解文書（項番 9,17）において、ランキング検索ではキーワードベース検索のほうが正解文書が上位にランクされたものの、絞り込みにより、意味検索のほうが正解文書が上位にランクされた。絞り込み効果の値を見ても、意味検索がキーワードベース検索に劣るものは 4 つのみ（項番 8,12,19,20）であるが、1 つ（項番 12）を除き、その差は僅差である。全般的に、キーワードベース検索よりも意味検索のほうが絞り込み効果が高いことがわかった。

6. 今後の展望と課題

今回は限られた範囲での実験しかできなかった。また、なぜ意味検索がキーワードベース検索よりも絞り込み効果が高いのか、その要因を分析するには至らなかった。今後は正解文書数を増やし、より客観的な評価をしていく他、実験結果の更に詳細な分析を行いたい。

参考文献

[1]大倉清司,潮田明(2012) 意味検索のプロトタイプシステムの構築. 言語処理学会第 18 回年次大会予稿集. 2012.

[2] Gerard Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw Hill Book Co., New York, 1983.

[3] Fillmore, Charles J. (1968) The Case for Case In: E. Bach and R.T. Harms (eds) Universals in. Linguistic Theory. Holt, Rinehart and Winston, New York. pp. 1-88

[4] 富士通.英日・日英翻訳ソフト ATLAS. http://software.fujitsu.com/jp/atlas/

Z：検索順位が100位以内

項番	課題番号	ｸｴﾘｰ	正解	意味検索			キーワードベース検索		
				検索順位(R)	絞り込み順位(B)	log(R/B)	検索順位(R)	絞り込み順位(B)	log(R/B)
1	1	1	1	12684	4444	0.455	***294***	***255***	0.062
2	1	1	2	887	771	0.061	***286***	***249***	0.060
3	2	1	1	***5035***	***2311***	0.338	7986	5176	0.188
4	2	2	1	16	10	0.204	700	617	0.055
5	3	1	1	1016	275	0.568	12	6	0.301
6	3	1	2	18201	1533	1.075	***816***	***643***	0.103
7	4	1	1	***1105***	***96***	1.061	4277	928	0.664
8	4	1	2	80	5	1.204	1298	66	1.294
9	4	1	3	10209	***505***	1.306	***5908***	1307	0.655
10	4	1	4	***2025***	***137***	1.170	15141	2357	0.808
11	4	2	1	***237***	***30***	0.898	4027	1119	0.556
12	4	2	2	4	2	0.301	1155	70	1.217
13	4	2	3	***2846***	***225***	1.102	6013	1561	0.586
14	4	2	4	***771***	***82***	0.973	12570	2258	0.746
15	5	1	1	7457	150	1.696	***125***	***23***	0.735
16	5	1	2	101085	570	2.249	29	11	0.421
17	5	2	1	1581	***269***	0.769	***782***	458	0.232
18	5	2	2	3647	419	0.940	87	82	0.026
19	6	1	1	***774***	***462***	0.224	1369	779	0.245
20	6	1	2	***154***	***108***	0.154	768	495	0.191
平均（全て）						**0.837**			**0.457**
平均（Zを除く）						**0.806**			**0.416**

図 3. .実験結果