2C3-05

意味構造検索システムの検索履歴を利用した検索ヒントの抽出

Hint Extraction from History of Semantic-Structure-Based IR Interaction

宮田 高志[*1]　橋田 浩一[*2*1]
Takashi Miyata　Kôiti Hasida

[*1]独立行政法人 科学技術振興機構, CREST
CREST, Japan Science and Technology Agency

[*2]独立行政法人 産業技術総合研究所 情報技術研究部門
ITRI, National Institute of Advanced Industrial Science and Technology

We are currently developing a information retrieval system that makes use of graph-matching rather than keyword-matching between a query and documents. The system analyzes documents in the database in advance and allows users to input graphs that represent semantic structure of contents to be retrieved. The system also allows users to manipulate the graph to revise their queries. This paper proposes a method to provide useful hints for revising users' queries by using histories recorded during interaction between the system and the other users.

1. はじめに

情報検索の分野では、統語的・意味的な構造を使った検索方法の提案やその効果に関する研究がこれまでにもいくつか報告されている [7, 4, 5, 8]。しかし結果はいずれも限定的であり、多くの研究者から「構造を使って検索を行なっても “性能” はほとんど向上せず、構造を解析・付与するコストには見合わない」とされてきた。

一方、近年の計算機性能の向上および統語解析における統計的アプローチの成功に伴って、構造を解析・付与するコストは大幅に下がっている。また、向上しないとされてきた “性能” は標準データと標準テストを使って測定された静的な性能 (例えば精度や再現率) であり、人間が検索システムとインタラクションしながら欲しい情報を得るような場合に重要となる性能 (例えば検索にかかる時間や問合せ改訂のためのヒントの有用性) とは必ずしも対応しない。

このような考察の下、我々は、従来のキーワード照合に基づく方法に代わって、問合せと文書の意味構造に対するグラフ照合に基づく検索システムを開発している。これまでに、意味構造を検索に用いることでユーザに対して問合せ改訂のために有用なヒントが提示できること、およびそのヒントを使って実際に検索の効率が向上することを示した [6]。

我々の検索システムでは、問合せグラフと文書グラフを (近似的に) 照合する際に、文書側の情報を利用することで問合せに応じた適切な類義語を提示する。原理的には、この方法を頂点や辺にも適用することは可能である。すなわち、その辺や頂点を追加した時に、現在ある程度適合している文書のスコアがどれだけ増加するかによって、類義語と同じように頂点や辺のスコアを定義することはできる。しかし、グラフの構造を変化させるということは質問の意味を変更するということであり、定義の妥当性に疑問が生じる。そこで本論文では、何人かのユーザが意味構造に基づく検索システムを使って同じような情報を検索した時の履歴から、その情報を検索する時に役に立ちそうなヒントを抽出する方法について考察する。

連絡先: 宮田 高志 (miyata.t@carc.aist.go.jp)

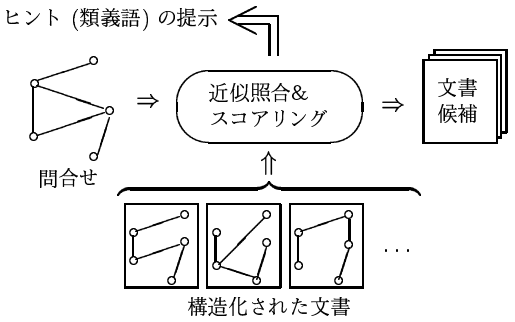

図 1: 意味構造に基づく検索システム

2. 意味構造に基づく検索

我々の意味構造に基づく検索システムの構成を図 1 に示す。システムは検索に先立ってデータベース中の全ての文書をあらかじめ統語解析しておき、その結果をもとに内容語を頂点、それらの間の係り受け関係を辺とするような**文書グラフ**に変換しておく。ユーザは問合せとして意味構造を表すグラフ (**問合せグラフ**) を入力し、システムは各文書グラフに対して最もよく合致する部分グラフを計算し、そのスコアの順に文書候補をソートしてユーザに提示する。ユーザは、問合せとしてグラフを直接入力するかわりに自然言語文を入力してシステムに解析させた結果を使うこともできる。

一般にはキーワードのリストよりもグラフの方が検索条件が詳しくなるので、一回で欲しい情報を含んだ文書が得られることはなく、ユーザは検索結果を見ながら類義語・関連語を追加して条件を緩和したり問合せの構造を修正したりして目的とする情報を含んだ文書を探すことになる。その結果、検索履歴としてユーザが入力した問合せグラフ G_h とそのグラフに対して施した修正操作 (の集合) op の対 (G_h, op) が記録される。G_h の頂点は類義語の集合、辺はそれらの間の依存関係を表す。なお、我々の検索システムでは、依存関係の種類および向きを無視している。これは、実装を簡単にするためと、問合せおよび文書を無向グラフとすることで検索条件を緩和するためという二つの理由からである。

3. 検索履歴からのヒントの抽出

検索履歴として得られた、問合せグラフとその修正操作の対の集合を使って新たな問合せグラフに対する修正操作のヒントを計算することが本論文の目的である。ここでは最も単純な方法として、k-最近傍法を用いることにする。すなわち、現在の問合せグラフ G_q と最も "近い" グラフを持つ k 個の対 (G_h, op) を検索履歴から探し、それぞれの G_h に対して行なわれた操作 op の中から最も頻度が高いものを選択して提示する。具体的な方法は、以下の通りである:

1. 問合せグラフ $G_q = (V_q, E_q)$ および履歴中のグラフ $G_h = (V_h, E_h)$ において、各頂点 $v \in V_q, V_h$ は類義語の集合、各辺 $e \in E_q, E_h$ はそれらの間の依存関係を表すとする。以下では頂点 $v \in V_q, V_h$ に対して、$W(v)$ と書いて v で指定されている類義語の集合を表すものとする。

2. G_q から G_h の上への写像 $f : V_q \to V_h \cup \{\perp\}$ に対して、近さの度合いを表すスコアを次のように定義する:

$$\begin{aligned} \mathrm{score}(f) &= \sum_{v \in V_q} \mathrm{sim}(W(v), W(f(v))) \\ &\quad - \sum_{(x,y) \in E_q} \mathrm{diff}(f(x), f(y)) \qquad (1) \\ \mathrm{sim}(X, Y) &= \frac{2|X \cap Y|}{|X| + |Y|} \qquad (2) \\ \mathrm{diff}(u, v) &= \begin{cases} d(u, v) - 1 & \text{if } u \neq v \\ 1 & \text{if } u = v \end{cases} \qquad (3) \end{aligned}$$

   ここで、$\mathrm{sim}(X, Y)$ は二つの類義語の集合 X, Y の間の近さ[*1]、$\mathrm{diff}(u, v)$ は G_q の辺と対応する G_h のパスとの差を表し、$d(u, v)$ は二つの頂点 u, v の間の G_h における最短距離である[*2]。

3. 検索履歴中の問合せグラフと操作の対 (G_h, op) を、G_h と現在の問合せ G_q とのスコアの最大値 $\max_f \mathrm{score}(f)$ の順にソートし、上位 k 個の対における操作 op を候補とする。この時、op 中で指定されている G_h 中の頂点を対応する G_q の頂点に読み換える。「辺の追加なのに端点に対応する点が G_q にない」というように読み換えが不可能な操作は候補には含めない。また n 通りの読み換えが可能な時は、読み換えた各操作の頻度を $1/n$ とする。

4. 候補となった操作が一つ以上あれば、その中で最も頻度が高いものをヒントとしてユーザに提示する。

図 2 に、上記のアルゴリズムを実装し、検索システムの評価実験 [6] において得られた検索履歴 (ユーザ 8 人、検索課題 12 題に対する 48 セッション分) に対してアルゴリズムを適用した例を示す。図では三番目の頂点 (i.e. 一番右端の頂点) に「少年」という類義語を追加することが提示されている。

4. 関連研究

Mitra ら [5] は Wall Street Journal や AP 通信など約 21 万文書を対象に、TREC の 50 の評価課題を使って、句を索引に使った時の効果について調べている。彼らは、キーワードだけで正解が上位にランクされるような場合は句を使っても精度はほとんど向上せず、無関係だが上位にランクされるような文書を排除する効果もなかったと報告している。その理由として、無関係な文書が上位にランクされるのは多く場合、キーワードだけの検索質問が曖昧であるためで、句を使ってもそれらの曖昧性の一つが強調されるだけで排除されるわけではないからだとしている。このことから彼らは、句に関する情報は下位にランクされた文書を再評価する時に使うべきであると結論付けている。日本語についても宮川ら [8] が毎日新聞約 43 万件を使った評価実験で同様の結論を得ている。

情報検索において、ユーザが検索システムのパラメータを直接もしくは間接的に調整することを一般に**関連性フィードバック**という。キーワードに基づく情報検索でよく使われる関連性フィードバックとしては、システムが提示した文書の関連度をユーザに評価してもらうという方法がある。この方法ではユーザが多くの文書を評価するほど、検索精度が高まる [1] が、関連があるかどうか注意して文書を読むというのはユーザにとって負担が大きい。

また Harabagiu ら [2] は、キーワード照合で文書から該当する部分を候補として抽出したあと、各候補をその場で解析し、統語的・意味的な構造を比較して最終的な出力を計算するという質問応答システムを開発した。彼女らのシステムでは、候補の数が多すぎたり少なすぎたりした時に自動的に類義語・関連語を追加したり、よく入力される問合せをその答とともにキャッシュしたりといった工夫もされている。

意味構造に基づく検索・関連性フィードバック・質問応答とも、ユーザの検索要求に関する情報をできるだけ多く引き出してより "よい検索" を行なおうという目的は同じであるが、検索システムを使うのが人間である以上、その "よさ" は人間を含めて評価すべきである。その意味で今後は TREC-6 interactive IR track [3] のような、詳細な実験計画に基づいた評価が重要となると思われる。

5. おわりに

意味構造に基づく検索システムにおいて、その検索履歴を使ってユーザにヒントを提示する方法を提案した。またこの方法を実際に実装し、評価実験で得られた検索履歴から適切なヒントが抽出できることを確かめた。

3. 節で示したアルゴリズムは、ヒントを提示することを目的としており履歴の内容だけを使っていた。しかし、どのユーザの履歴かということまで記録しておけば、興味が似ている他のユーザを紹介するということも (プライバシーに配慮する必要はあるが) 原理的には可能である。このようなシステムはオンラインショッピングなどではすでに実用化されているが、各ユーザから情報を得るところがボトルネックになることが多い。これは、ユーザからの情報が乏しければシステムが提供するサービスの質も向上せず、サービスの質がそこそこであればユーザもそれなりの情報しか入力しなくなる、という負のフィードバックが働くからである。

グラフを直接編集するというインターフェースは改良の余地があるが、意味構造を入力することで検索の効率が向上したり、興味が似ている他人を高精度で見つけ出したりできることがユーザに広く理解されれば、上記のボトルネックが解消される可能性が高い。

---

*1 ただし、$W(\perp) = \phi$ とする。また、類義語の集合の間の近さとして Dice 係数を用いているが、とくに強い根拠はない。

*2 検索においても (1) と同様のスコアを用いている。

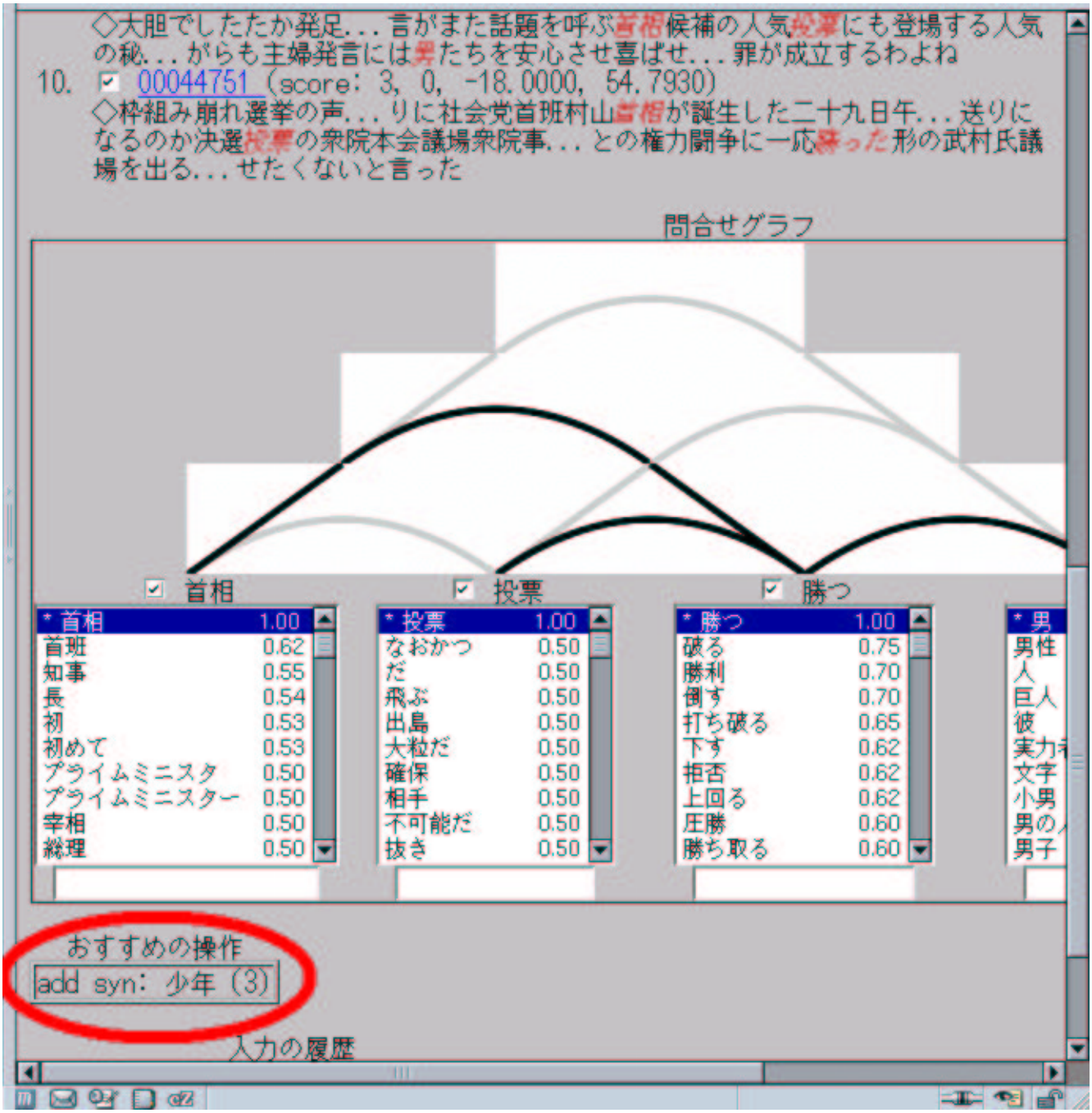

図 2: 提示されたヒントの例

参考文献

[1] Chris Buckley, Gerard Salton, and James Allan. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 7th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 292–300, 1994.

[2] Sanda Harabagiu, Dan Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan Bunescu, Roxana Gîrju, Vasile Rus, and Paul Morărescu. The role of lexico-semantic feedback in open-domain textual question-answering. In *Association for Computational Linguistics*, pp. 274–281, France, July 2001.

[3] Eric Lagergren and Paul Over. Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 164–172, Australia, 1998.

[4] Geoffrey Z. Liu. Semantic vector space model: Implementation and evaluation. *American Society for Information Science*, Vol. 48, No. 5, pp. 395–417, 1997.

[5] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactical phrases. In *RIAO '97*, pp. 200–214, 1997.

[6] Takashi Miyata and Kôiti Hasida. Information retrieval based on semantic structures. In *Proceedings of the 2nd Language and Technology Conference*, pp. 167–171, Poznań, Poland, April 2005.

[7] Tomek Strzalkowski. Natural language information retrieval. *Information Processing and Management*, Vol. 31, No. 3, pp. 397–417, 1995.

[8] 宮川和, 徳永健伸, 田中穂積. 格フレームを用いた情報検索. 第四回年次大会発表論文集, pp. 112–115, 九州大学, March 1998. 言語処理学会.