

キーワード検索における検索結果の拡張

田邊 翼[†] 清水 敏之^{††} 吉川 正俊^{††}

[†] 京都大学工学部情報学科 〒606-8501 京都市左京区吉田本町

^{††} 京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町

E-mail: [†]tanabe@db.soc.i.kyoto-u.ac.jp, ^{††}{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし 本稿では、用いられた問合せ語を基に検索対象のデータからその問合せ語に関連する語を抽出し、検索結果を拡張する手法の提案を行う。利用者が検索結果を網羅的に閲覧したい場合や検索結果を深く理解したい場合などに、検索結果の拡張は有効であると思われる。従来の内容や履歴に着目した拡張手法とは異なり、我々は問合せ語を基に問合せ語と対応するデータの関連や量に着目し問合せ語の拡張を行い、検索結果を徐々に拡張して利用者に提示する。我々の手法では抽出された関連する語を用い、どのような結果が拡張されたかを利用者に提示することができ、本手法を用いることによって、利用者は検索結果に対する理解を深めることができる。我々は提案した手法の実装を行い、実験において論文情報をタイトルから検索するような状況における検索結果の拡張事例を挙げて有効性を議論した。

キーワード 検索結果の拡張, 問合せ語, キーワード検索

An Expansion Method of Keyword Search Results

Tsubasa TANABE[†], Toshiyuki SHIMIZU^{††}, and Masatoshi YOSHIKAWA^{††}

[†] Undergraduate School of Informatics and Mathematical Science, Faculty of Engineering, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, 606-8501 Japan

^{††} Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, 606-8501 Japan

E-mail: [†]tanabe@db.soc.i.kyoto-u.ac.jp, ^{††}{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

1. はじめに

近年、計算機の普及に伴い日常生活の様々な場面で計算機が多くの人々に使われるようになった。それに伴い、計算機を用いることによって情報を管理し、その管理した情報の中から必要な情報を検索する機会が増えてきた。

検索にも様々な種類があり、提示されたカテゴリの中から利用者が興味を持った語や、知りたいと思った語を選んでいくことによって、求める情報を検索することのできるカテゴリー検索などが考えられる中、我々はキーワード検索というものに着目した。キーワード検索は、対象となるデータセットから利用者が問合せに用いた問合せ語を含むデータのみを取得する手法である。キーワード検索は求める対象が明確な場合に特に有効であり、予備知識を必要とせず、直接問合せ語を指定して求めたい情報を検索することができ、一般に複数の問合せ語を指定することもできる。

我々は検索結果から得られた情報を基に検索結果を拡張する手法を提案する。一般的なキーワード検索では対象データの中

から問合せに関連するデータの絞り込みを行うのみであるが、例えば検索結果を絞り込んだ後に、検索結果と関連はあるが検索結果には含まれないデータを取得したいという要求があると思われる。我々が提案する手法によって、検索結果を絞り込んだ後に、その結果を別の観点から広げて検索結果を拡張することができ、幅広い視点を利用者に与えることができる。

我々は、キーワード検索において、利用者が適切に問合せ語を作成できている場合と作成できていない場合の両方に関して議論を行う。それぞれの場合に関して我々は異なった手法で検索結果の拡張を行い、その両方に関して検索結果の拡張は有効だと考えている。利用者は文書集合を問合せ語によって絞り込むことが可能だが、適切な問合せ語を作成できていない場合は求める情報が得られない場合もある。そのような場合、検索結果に関連のある部分を拡張して利用者に提示することによって、求める情報を手に入れることができる可能性がある。また、たとえ問合せ語を適切に作成できている場合であっても、得られた検索結果に関連した情報を拡張して提示することによって幅広い視点を利用者に提供することができ、その結果として利用

者は理解を深めることができる。

キーワードによる検索条件の指定には、利用者が問合せ語を適切に作成できていない場合や、検索対象が曖昧で問合せ語を用いて直接検索結果を絞り込むのが困難な場合においては、検索結果を拡張する以前に問合せ語の拡張を行うことにより利用者に対する支援を行うことができると思われる。問合せ語の拡張とは、利用者が用いた問合せ語にさらにそれとは異なる問合せ語を追加したり、問合せ語の組合せを変更したりする手法である。しかし、先に述べたように、直接対象を指定できない状況に対しても、検索結果に関連する部分を拡張して利用者に提示することによって、問合せ語の拡張とは異なった形で理解支援ができる。

利用者は検索する際に用いた問合せ語に関して興味があると思われるため、用いられた問合せ語を基に、検索対象のデータからその問合せ語に関連する拡張問合せ語を抽出し、問合せ語の拡張を行う。ここで、拡張問合せ語とは問合せ語の拡張に用いる問合せ語のことである。そして、拡張問合せ語に基づいて検索結果の拡張を行う。拡張する部分は様々なものがあり、それらの部分についてランキングを行い、その拡張した検索結果を利用者に提示し、利用者はその提示された検索結果の中から興味のある情報を選択する。その際に、拡張した検索結果を返すために用いた拡張した問合せ語を利用者に提示することによって、利用者はどのように拡張した検索結果が提示されたのかということを知ることができ、理解を深めることができる。

また、拡張された結果をさらに拡張することによって、利用者は興味のある情報を増分的に拡張していくことが可能であり、そうすることによって利用者はより幅広い知識、視点を得ることができる。我々が提案する手法は、利用者が検索結果を網羅的に閲覧したい場合や、検索結果を深く理解したい場合などに有効であり、増分的に検索結果を拡張することによって利用者の知らなかった、あるいは思いつかなかった知識の提示も行いうることができる可能性がある。

例えば文献データベースに対して、利用者が“XML retrieval”という問合せで検索を行い、その検索結果として“XML retrieval”に関する文献集合が得られたとする。我々の手法では、まずその検索結果内のデータから拡張問合せ語を抽出する。この例においては、例えば“hyrex”, “inex”, “xtree”, “ranked”, “information”などといった拡張問合せ語が抽出される。それらの語を用いて問合せ語を拡張し、拡張する検索結果のスコアを算出しランキングを行う。その結果として、“xtree XML not(retrieval)”という拡張した問合せによって“xtree”や“XML”という問合せ語は含むが、元の問合せに含まれていた“retrieval”という問合せ語を含まない“XTree for Declarative XML Querying”といったタイトルの文献や、“db/ir not(XML) retrieval”という拡張した問合せによって“Internet-accessible information retrieval tools for advanced DB/IR courses”といったタイトルの文献を拡張した検索結果として利用者に提示する。また、利用者は拡張された結果を選択することができ、その選択された拡張検索結果のデータも用いることによって、さらに拡張を繰り返し行っていくことがで

きる。なお、我々は本論文においては対象データをプレーンテキストとして扱い、問合せも単純なキーワード集合として扱う。

2. 関連研究

我々は検索結果の拡張を提案しており、その検索結果の拡張のために問合せ語の拡張を行っている。問合せ語の拡張は、例えば検索結果の量をより少なくするために絞り込むものであったり、代替問合せとなる問合せ語を返すものである。ここで、問合せ語の拡張に関する研究として、堀ら [1] は Web ページの検索において、異なる目的で入力された問合せ語が混在するとよい結果が得られないという問題に着目し、利用者の閲覧履歴を基に問合せ語の拡張を行う手法を提案している。また、岡部ら [2] は利用者が一定数の文書の適合性を判定することで、その判定情報を利用して問合せ語の拡張を行う手動フィードバック型の方法に着目し、利用者が適合文書の一つを見つけるまでの文書判定情報のみを利用した場合に有効な問合せ語の拡張方法を提案している。

それに対し、検索結果の拡張は、検索結果をある程度絞り込んだ後に、その検索結果を別の観点から拡張することによって、幅広い視点を利用者を与えることができる。しかし、問合せ語に基づいて検索結果の拡張を行うためには、利用者が用いた問合せ語以外の語も抽出しなくてはならないため、その意味では問合せ語の拡張を行っているともいえる。

検索結果の拡張に関する研究としては、佐藤 [3] や増井ら [4] によるものなどがある。佐藤 [3] は履歴に着目して検索結果の拡張を行っており、個人の情報を検索、閲覧するシステムを実現している。佐藤は Web ページの閲覧履歴などを検索対象のデータとして扱っており、そのデータの時間情報を用いることで、ある一定時間以上ページを閲覧していたかなどの情報を基に拡張を行っている。また、増井ら [4] による、あるファイルに対して保存場所が近いといった情報や、類似度を計算することによって算出した内容が近いといった情報を連想的に検索することによって、求めたい情報を探し出す手法のように、内容や履歴に着目したデスクトップ検索に関する研究もなされている。

それに対して我々はキーワードに着目しており、従来の履歴を用いた検索結果の拡張手法や、内容比較に基づいた検索結果の拡張手法とは異なり、問合せ語を基に、問合せ語と対応するデータの関連や量に着目し、検索結果を徐々に拡張して利用者に提示する。

我々が提案する検索結果の拡張は検索結果の理解支援の一種であると考えられるが、検索結果の理解支援に関する研究としては様々なものがある。例えば Ziyang ら [5] は、検索結果の特徴を示す情報量を付加する手法を提案している。また、Quoc ら [6] は、ある問合せ語を用いることによって返ってきた検索結果と同じ検索結果を返す元々の問合せ語とは異なった問合せ語 (instance-equivalent queries) を返す手法を提案している。

3. 問合せ語の拡張

我々は利用者が検索時に利用した問合せ語から得られた結果

を拡張する。なお、我々はデータセットや問合せ語に関しては構造化されていないものを対象としている。

我々はキーワードに基づいた検索結果の拡張を提案しており、検索結果の拡張に関する既存の研究とはそもそも着目点が異なっているが、佐藤 [3] や増井ら [4] のような既存の研究ではデスクトップサーチとの連携を目的としているものがほとんどである。それに対して我々の手法は、ランキングを考慮することによって妥当と思われる結果を拡張することができ、検索結果を徐々に拡張して利用者に提示する。

例えば、ある利用者が DBLP^(注1) のような文献データベースにおいて、“XML retrieval” という問合せで検索を行うことによって、“XML retrieval” に関する文献集合が得られるが、その後我々は検索結果の拡張を行う。ここで、検索結果を拡張するためにまず問合せ語の拡張を行う。“XML” や “retrieval” といった問合せ語に関連のある拡張問合せ語を検索結果として得られた文献集合内のデータから抽出する。そして、その抽出された拡張問合せ語を用いて問合せ語を拡張し、例えば拡張問合せ語が “xtree” であった場合、“xtree XML not(retrieval)” というような拡張された問合せと共に、その拡張した問合せによって返ってくる拡張した検索結果を提示する。

3.1 問合せ語の拡張手法

まず問合せ語の拡張について、堀ら [1] や Joaquin ら [7] のような既存の研究の手法を用いるという選択肢もあるが、既存の研究では検索結果をさらに絞り込むための拡張や、元々の問合せ語と似た意味を持つ問合せ語を類似度の算出などによって拡張を行っている。しかし、問合せ語だけからの拡張では元々の検索結果との関連が保証されないために、今ある検索結果を網羅的に閲覧するといったような場合にはあまり効果的ではない。

検索結果を拡張するにあたって重要な点は、拡張した部分が利用者が元々用いた問合せ語や検索結果とどのように関わっているのかということである。つまり、例えば先ほどの “XML retrieval” という問合せで検索した例において、その拡張問合せ語が “information” である場合と “xtree” である場合とでは意味合いは大きく異なる。DBLP のような文献データベースにおいて、“information” は、様々な語との関わりが強い語であり、利用者が用いた問合せ語だけでなく他のキーワード集合とも強い関連がある。それに対して、“xtree” は利用者が用いた問合せ語以外にはあまり関連がない語であり、“information” と比較するとより利用者の用いた問合せ語に特に強く関連がある語であるといえる。つまり、拡張問合せ語自体で検索を行った際に、返ってくる検索結果量が多いものほど様々な語との関連が強い傾向にあり、返ってくる検索結果量が少ないものほど利用者が用いた問合せ語以外にあまり関連がない傾向にある。

利用者が生成した問合せ語は、その利用者にとって検索時点では一番興味のあった語であると考えられ、その問合せ語によって得られる情報を用いることによって問合せ語を拡張することが妥当であると思われる。つまり、問合せ語の拡張に関し

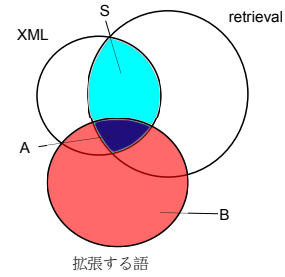


図 1 問合せ語の拡張

ては、利用者が検索時に利用した問合せ語から得られた検索結果内の語を用いる。

図 1 に示すように、利用者が検索時に利用した問合せ語から得られた検索結果を S とし、その検索結果量を $|S|$ とする。その問合せ語に、検索結果内の語から得られた拡張問合せ語を拡張することによって、作成された新たな問合せを用いて得られた検索結果を A とし、その検索結果量を $|A|$ とする。拡張問合せ語自体で検索することによって得られた検索結果を B とし、その検索結果量を $|B|$ とする。我々は検索結果の拡張にあたり、(1) 式、(2) 式、(3) 式によって求まる値を利用することとした。

$$E_S = \log \frac{1}{\frac{|S| - |A|}{|S|}} \quad (1)$$

$$E_B = \log \frac{1}{\frac{|B| - |A|}{|B|}} \quad (2)$$

$$\delta_{exp} = \begin{cases} 0 & (|B| - |A| = 0) \\ 1 & (otherwise) \end{cases} \quad (3)$$

まず、(1) 式の E_S について、この部分では、 $|S|$ がわかっているときの $|S| - |A|$ の情報量を求めており、拡張した問合せによって返ってきた検索結果量が、元々の検索結果量に対してどの程度占めているかを計算している。つまり、その拡張問合せ語が元々の検索結果に対してどの程度の関連があるのかを求めており、値は大きければ大きいほど拡張問合せ語と元々の問合せ語の関連が強いといえる。この部分に関しては、例えば “information” や “ranked” といったデータベース全体として出現頻度の高い拡張問合せ語ほど、値が大きくなる傾向にある。

次に、(2) 式の E_B について、この部分では、 $|B|$ がわかっているときの $|B| - |A|$ の情報量を求めており、拡張問合せ語が元々の問合せ語の結果にどの程度関連があるのかを求めている。拡張問合せ語自体の検索結果量が多いということはすなわち、その語が他のキーワード集合とも関連がある可能性が高く、“XML retrieval” という問合せで検索した例においては、拡張問合せ語が “information” の場合がそれに当てはまる。“information” という語は用いられた問合せ語だけでなく、他の様々なキーワード集合にも関連があるために、用いられた問合せ語のみと関係が強い訳ではないと思われる。つまり、このような他のキーワード集合とも強い関連があるような語ほど (2) 式の E_B の値は小さくなる傾向にある。利用者が適切に問合せ語を作成できていない場合、拡張して提示する問合せ語や検索結果としてはごく限られた結果を返すものより、幅広く検

(注1) : <http://www.informatik.uni-trier.de/~ley/db/>

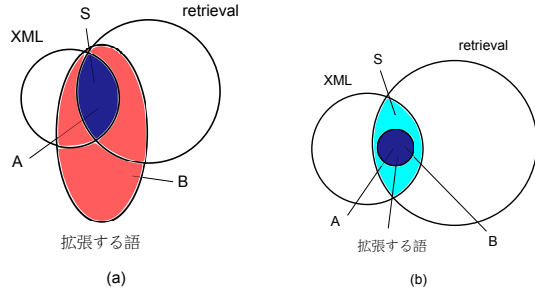


図 2 問合せ語の拡張における特殊な例

索結果を返すものの方が効果的であり、そのような場合は (2) 式の E_B の値が小さいものほど効果的に働く。

我々は利用者に検索結果を拡張して提示するが、拡張された検索結果を実際に閲覧するのは利用者であり、様々な情報をもつ結果を一度に拡張するのではなく、徐々に拡張して利用者に提示することで、利用者の負担を軽減できると考える。“xtree”のような拡張問合せ語は、利用者が“XML retrieval”という問合せで検索した例においては、利用者が用いた問合せ語以外にあまり関連がない拡張問合せ語であり、検索結果を徐々に拡張して提示するための拡張問合せ語として適切である。つまり、(2) 式の E_B の計算式では、利用者が問合せ語を適切に作成できている場合に効果的な拡張問合せ語ほど値が大きくなると考えられる。

最後に、(3) 式の δ_{exp} は、後に行う検索結果の拡張に関係する部分であり、問合せ語を拡張するにあたって、検索結果の拡張を行うことができるか否かを調べている。仮に図 2(b) のように拡張する部分が存在しなかった場合、その選出された語で問合せ語を拡張する必要性はなくなるため、スコアが 0 になる。拡張する部分が存在する場合は、その選出された語で問合せ語を拡張することは有効であり、拡張の候補として挙げることができる。このように、問合せ語を拡張することによって検索結果の拡張も行うことができるか否かを調べているのが (3) 式の δ_{exp} である。

図 2(b) のように拡張する部分が存在しないものは特殊な場合であり、それとは別に図 2(a) のように元々の検索結果として得られた部分全てが、拡張問合せ語自体で検索した場合に返ってくる検索結果の中に含まれる特殊な場合もある。図 2(a) のような場合においては、そのような特殊な状態の拡張問合せ語であることを示し、後に行う検索結果の拡張において、図 2(a) や 2(b) のような特殊な場合には当てはまらない拡張問合せ語とは分けて提示し、検索結果を拡張する。

さて、検索結果を拡張するに当たって、以下のように状況を分類した。

$$\left\{ \begin{array}{l} \text{利用者が問合せ語を} \left\{ \begin{array}{l} \text{適切に作成できている} \\ \text{適切に作成できていない} \end{array} \right. \\ \text{利用者が検索結果の拡張量を} \left\{ \begin{array}{l} \text{多くしたい} \\ \text{少なくしたい} \end{array} \right. \end{array} \right.$$

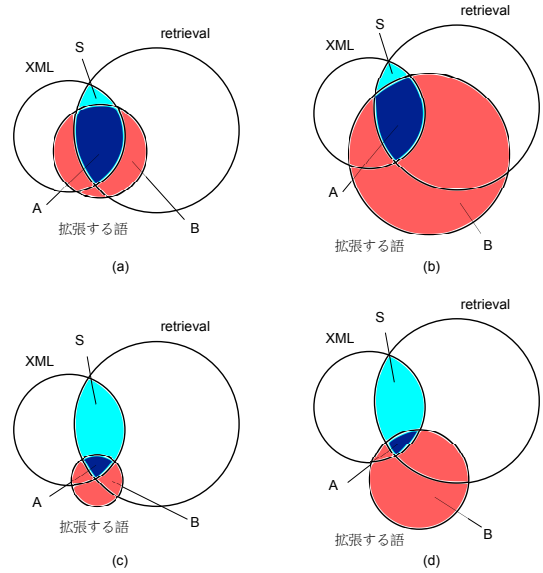


図 3 問合せ語の拡張

これらは、先に述べた (1) 式の E_S と (2) 式の E_B の組合せによって表すことができる。つまり、利用者が問合せ語を適切に作成できている場合は、 E_S の値が大きく、検索結果の拡張量を多くしたい場合は E_B の値が小さいものほど優先されるべきである。このように、我々はそれぞれの状況に対応して以下の (4) 式、(5) 式、(6) 式、(7) 式により拡張問合せ語のスコアを算出することを提案する。

$$score_{query}^{(1)} = (E_S + E_B) \cdot \delta_{exp} \quad (4)$$

$$score_{query}^{(2)} = (E_S - E_B) \cdot \delta_{exp} \quad (5)$$

$$score_{query}^{(3)} = (-E_S + E_B) \cdot \delta_{exp} \quad (6)$$

$$score_{query}^{(4)} = (-E_S - E_B) \cdot \delta_{exp} \quad (7)$$

利用者が適切に問合せ語を作成できている場合は、図 3(a) や図 3(b) のように E_S の値が大きく、検索結果と関連の強い拡張問合せ語を用いることが有効であると思われる。その中で、利用者が自分の負担を考慮し、拡張量を小さくしたい場合は図 3(a) のように E_B の値が大きい拡張問合せ語を選択することが有効であり、(4) 式を用いてスコアを算出することが妥当である。一方、拡張量を大きくして多くの検索結果を閲覧したいと考える場合は、図 3(b) のように E_B の値が小さい拡張問合せ語を選択することが有効であり、(5) 式を用いてスコアを算出することが妥当である。

利用者が適切に問合せ語を作成できていない場合は、図 3(c) や図 3(d) のように E_S の値が小さく、別の視点を提供できるような拡張問合せ語を用いることが有効であると思われる。さらに利用者の負担に応じて、図 3(c) のように (6) 式によるスコアが大きくなるような語を選択したり、図 3(d) のように (7) 式によるスコアが大きくなる語を選択したりすることが考えられる。

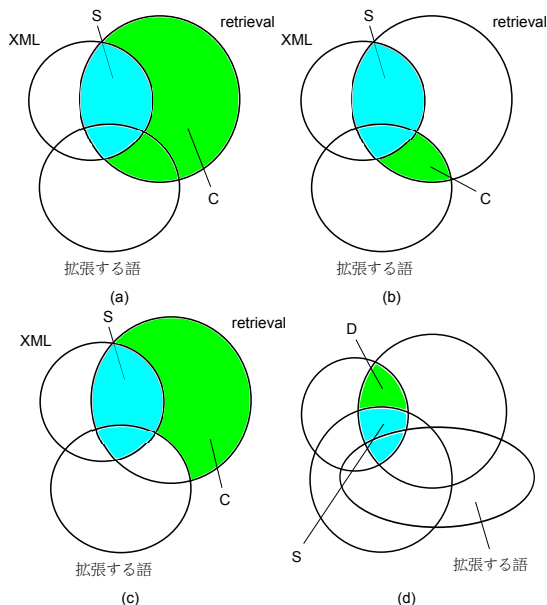


図 4 検索結果の拡張部分

(4) 式, (5) 式, (6) 式, (7) 式のどれを用いて問合せ語の拡張を行うべきかは利用者依存であるために, どの場合を想定して拡張を行うのかを利用者が選び, その選択された場合に従って問合せ語の拡張を行うことによって, 利用者の意図に従った拡張が行える。

また, ある閾値以上のスコアを返した語についてのみ検索結果の拡張を考えることによって, この後に行う検索結果の拡張における計算量を削減することが考えられる。

4. 検索結果の拡張

4.1 検索結果の拡張部分

拡張された問合せ語を用いることによって, 検索結果を拡張する。その際に拡張される検索結果はランキングして提示し, 拡張された検索結果を返す拡張した問合せ語の提示も行う。例えば, 利用者が先に述べた “XML retrieval” という二つの問合せ語で検索を行った場合, 図 4(a), 図 4(b), 図 4(c) の S の部分が検索結果として返ってくる。その際に, 例えば図 4(a), 図 4(b), 図 4(c) の C の部分の検索結果を拡張して返すことが考えられる。これらの例からもわかるように, 拡張できる検索結果については様々な部分がある。そこで, どの部分を優先して返すのかを考える必要がある。

まず, 検索結果の拡張において, 検索結果を拡張する際に図 4(b) のように, 例えば問合せ語が “not(XML) retrieval xtree” といったように拡張問合せ語を用いる場合と, 図 4(a) のように, 例えば問合せ語が “not(XML) retrieval” といったように拡張問合せ語を用いない場合とに大きく分けることができる。用いる問合せ語の数が増え, 3 語で検索した場合, 図 4(d) のように利用者が作成した問合せ語によって返ってきた検索結果が S であったとすると, D の部分のように拡張問合せ語とは関連がない部分も出てくることもある。まず, 拡張問合せ語の利用によって拡張する部分を分割するか否かに関して考える。

我々の手法は, 利用者が検索を行うことで得られる情報を基

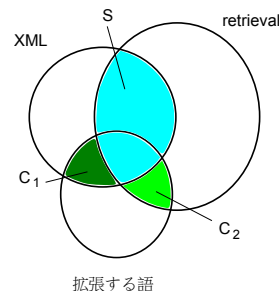


図 5 検索結果の拡張

に問合せ語を拡張し, 増分的に検索結果を拡張する。つまり, 繰り返し検索結果の拡張を行っていくことを想定している。それ故に, 仮に図 4(b) のように “not(XML) retrieval xtree” という問合せによる検索結果を拡張して利用者に提示したとしても, さらに検索結果の拡張を行い, 今度は図 4(c) のように “not(XML) retrieval not(xtree)” という問合せによって検索結果を拡張することによって, 結果的に図 4(a) の “not(XML) retrieval” という部分を全て閲覧することができる。それ故に, 拡張する検索結果としては, 拡張問合せ語を用いている図 4(b) における C の部分や図 4(c) における C の部分を優先して返すことに問題が無い。つまり, 後で増分的に検索結果を拡張できることを考えると, 検索結果を分割できる場合は, より小さな部分を優先して返すことが有効だと考えた。

次に, 図 4(d) の D の部分のように拡張問合せ語とは関連がない部分について, この部分についてはこの段階ではこれ以上細かく検索結果を分割することができないので, そのまま拡張する部分の候補として用いる。

4.2 検索結果の拡張手法

まず, 利用者が拡張量を小さくしたい場合について考える。拡張量を小さくしたいことから, 拡張量が小さいものを優先して返す。また, 拡張量を小さくしたい場合, 拡張量の大きい “retrieval” のような様々なキーワード集合と関連のあるものよりも, 拡張量の小さい “XML” のようなより特徴的な問合せ語を拡張した問合せに含んでいるものを優先して返す方が有効であると考えた。

拡張する部分の検索結果量を考慮する際に, 例えば図 5 のように拡張する部分 C_1 と C_2 の検索結果量がほとんど同じ場合についても考える。図 5 の C_1 と C_2 の部分についてどちらを優先して拡張して返すのか, 拡張する部分に関連のある語について検討し, その関連のある語の idf 値の平均の値を算出して利用することが有効であると考えた。例えば, 拡張する部分が図 5 の C_1 の部分であった場合, その拡張する部分に関連のある語は “XML” と拡張問合せ語であり, idf 値の平均とは “XML” の idf 値と拡張問合せ語の idf 値のそれぞれの平均を取ったものである。どの部分を拡張して返すのかということに関して, “XML” と “retrieval” 自体の検索結果量を比べると, 今回扱っている DBLP のような文献データベースの場合, 圧倒的に “retrieval” の方が多く, 検索結果からもより様々なキーワード集合と関連がある問合せ語であることが容易にわかる。

それ故に, 例えば拡張問合せ語が “rank” であった場合, “XML

rank”という問合せよりも“retrieval rank”という問合せによる検索結果の方が様々なキーワード集合と関連がある結果を返すことが考えられる。

ここで、拡張する部分に関連のある語の idf 値の平均の値を idf_{avg} とすると、例えば図 5 の C_1 部分の idf_{avg} に関しては、“XML” の idf 値と拡張する語の idf 値の平均を計算する。全文書数を $|D|$ とし、“XML” という問合せ語によって返ってきた検索結果を D_{XML} とし、その検索結果量を $|D_{XML}|$ とする。さらに、拡張問合せ語自体によって返ってきた検索結果を D_{exp} とし、その検索結果量を $|D_{exp}|$ とすると、 C_1 に対する idf_{avg} は、

$$idf_{avg} = \frac{\log \frac{|D|}{|D_{XML}|} + \log \frac{|D|}{|D_{exp}|}}{2} \quad (8)$$

となる。

つまり、関連のある語 W_i について、 W_i という問合せ語によって返ってきた検索結果を D_i 、その検索結果量を $|D_i|$ とし、関連のある語の総数を N 、全文書数を $|D|$ とすると、

$$idf_{avg} = \frac{\sum_{i=1}^N \log \frac{|D|}{|D_i|}}{N} \quad (9)$$

となる。

ここで、拡張する部分の検索結果を C とし、その検索結果量を $|C|$ とする。また、拡張する部分に関連のある語の idf 値の平均を idf_{avg} とし、その値を (9) 式で定義したものとすると、

$$score_{result} = \begin{cases} 0 & (|C| = 0 \text{ の場合}) \\ \frac{1}{|C|} \cdot idf_{avg} \cdot \delta_{query} \cdot \delta_{include_q} & (\text{それ以外}) \end{cases} \quad (10)$$

$$\delta_{query} = \begin{cases} 0 & (\text{拡張問合せ語のみを用いる場合}) \\ 1 & (\text{それ以外}) \end{cases} \quad (11)$$

$$\delta_{include_q} = \begin{cases} 0 & (\text{拡張問合せ語が含まれない場合}) \\ 1 & (\text{それ以外}) \end{cases} \quad (12)$$

が最大となるものを優先して拡張する。

(11) 式の δ_{query} について、例えば図 5 において拡張問合せ語が“xtree”であった場合、“not(XML) not(retrieval) xtree”という問合せで拡張されたもののスコアを 0 にしている。まず、利用者が適切に問合せ語を作成できている場合、その問合せ語が一つも含まれないようなものはあまり有用ではないと考えたためである。また、利用者が適切に問合せ語を作成できていない場合も同様で、作成した問合せ語の全てが不適切であることはまれであり、利用者が用いた問合せ語の一つも含んでいないような拡張問合せは有用でないと考えた。

次に (12) 式の $\delta_{include_q}$ について、例えば図 5 において拡張問合せ語が“xtree”であった場合、“XML not(retrieval) not(xtree)”といったように、拡張問合せ語が拡張した問合せに含まれないもののスコアを 0 にしている。まず、利用者が適切に問合せ語を作成できている場合、利用者は適切に問合せ語を作成できているために、“XML not(retrieval) not(xtree)”

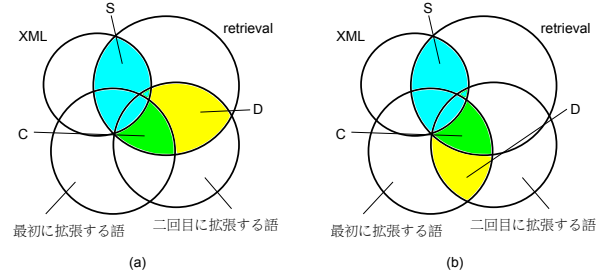


図 6 増分的な検索結果の拡張

という問合せで返される検索結果部分について閲覧したい場合は、“XML not(retrieval)”という問合せであらかじめ検索することが可能である。また、問合せ語の数が増えるにつれ、拡張問合せ語とは関係のない部分が増えていくことが予想され、そのような部分について閲覧したい場合は、あらかじめ問合せ語を指定して検索をすることができる。それ故に、拡張問合せ語が含まれないものは有用ではないと考えた。利用者が適切に問合せ語を作成できていない場合に関しても、適切に問合せ語を作成できていないが故に利用者としては拡張問合せ語の情報を閲覧したいと思われ、拡張問合せ語が含まれないようなものは有用ではないと考えた。

次に、利用者が拡張量を大きくしたい場合について考える。この場合は拡張する量の大きい“retrieval”のようなより様々なキーワード集合と関連のある問合せ語を問合せに含んでいるものを優先して返す方が有効であると考え、(10) 式の値が小さいものを優先して返す。つまり、 idf_{avg} 、 δ_{query} 、 $\delta_{include_q}$ のそれぞれの値を (9) 式、(11) 式、(12) 式で定義したものとすると、

$$score_{result} = |C| \cdot \frac{1}{idf_{avg}} \cdot \delta_{query} \cdot \delta_{include_q} \quad (13)$$

が最大となるものを優先して拡張する。

4.3 増分的な検索結果の拡張

我々の手法では、利用者は検索結果の拡張を一度だけではなく、増分的に興味のある情報を拡張していくことが可能である。図 6 に関して、例えば利用者が“XML retrieval”という問合せで検索を行った場合、S の部分が結果として返ってくる。ここで、利用者が検索結果の拡張を行い、さらに C の部分が拡張されて返ってきたとする。この状態からさらに拡張する際に、S の部分のみや C の部分のみに着目して検索結果を拡張するのではなく、S と C の両方を考慮することによって検索結果を拡張する。そうすることによって、(1) 式における $|S|$ の部分が変化し、例えば新たに図 6(a) や図 6(b) の D の部分を拡張することが考えられる。

今の例においては、まず“XML retrieval”という問合せで検索を行った場合、その問合せによって得られた検索結果が図 6 の S の部分に相当し、その検索結果内のデータから問合せ語の拡張に用いる拡張問合せ語を選出し、その拡張問合せ語を用いることによって問合せ語を拡張する。そして、拡張した問合せ語から検索結果を拡張する部分についてのランキングを行い、利用者に拡張した結果を提示することによって利用者が興味のある検索結果を選ぶことができる。まず、最初に検索結果の拡

表 1 “XML retrieval” における拡張問合せ語の候補

拡張問合せ語	E_S	E_B	$score_{query}^{(1)}$
hyrex	0.0081633	1.0986	1.1068
db/ir	0.0040733	0.69315	0.69722
eprum	0.0040733	0.69315	0.69722
bm25f	0.0040733	0.69315	0.69722
xtree	0.0040733	0.69315	0.69722

張を行った結果として、利用者が拡張された検索結果の中から図 6 の C の部分を選択したとする。さて、ここからさらにもう一度検索結果の拡張を行う場合、先ほど選択した C の部分の検索結果のデータも用いることによって、拡張問合せ語のスコアを算出する。その際に用いる式としては、(4) 式、(5) 式、(6) 式、(7) 式の四つが考えられるが、今は利用者が (4) 式を用いることを選択しているものとする。図 6 の検索結果 S の検索結果量を $|S|$ とし、検索結果 C の検索結果量を $|C|$ とすると、(1) 式の $|S|$ の部分が図 6 の $|S| + |C|$ に置き換えられる。よって、図 6 において、さらに拡張問合せ語を用いて作成された新たな問合せによる検索結果を A とし、その検索結果量を $|A|$ とすると、(4) 式は

$$score_{query} = \left(\log \frac{1}{\frac{|S|+|C|-|A|}{|S|+|C|}} + \log \frac{1}{\frac{|B|-|A|}{|B|}} \right) \cdot \delta_{exp} \quad (14)$$

となる。繰り返し拡張していくことより、最初に利用者が閲覧した検索結果から徐々に視点を変更していくことが可能となり、より利用者の望む情報を拡張して提供できると思われる。

また、利用者が適切に問合せ語を作成できていない場合、まず検索結果を閲覧し、求めたい情報がなかった場合は我々の手法によって拡張した検索結果からより利用者が欲しいものを選択することができる。そして、繰り返し検索結果の拡張を行うことによって検索結果を閲覧する視点が徐々に移っていくことから、欲しい情報を利用者に提供することを補助できる。

5. 実 験

本手法によってどのような語および結果が拡張されるのか、調査を行った。データセットとしては DBLP の文献データベースを使用しており、2010 年 11 月 29 日に公開されていたものを使用している^(注2)。論文数は 900557 件であり、今回はその中でもタイトルのみを用いて実験を行った。

まず、先に述べたように、例えば利用者が XML の検索についての論文を調べたいような状況だった場合に、“XML retrieval” という問合せで検索することを想定する。利用者が適切に問合せ語を作成できている場合に、拡張問合せ語のスコア式として (4) 式を用いることを選択した場合、拡張問合せ語の候補上位 5 位までについては表 1 のものが得られた。

利用者が適切に問合せ語を作成できている場合に、より網羅的に検索結果を閲覧することができる “xtree” や “eprum” といった問合せ語が選出できている。

さて、まず利用者が問合せ語を適切に作成できている場合に、

表 2 “XML retrieval” における拡張した検索結果を返す問合せ

問合せ	$score_{result}$
“XML not(retrieval) xtree”	9.1036
“XML not(retrieval) dataweb”	8.9009
“not(XML) retrieval db/ir”	8.7618
“not(XML) retrieval bm25f”	8.7618
“XML not(retrieval) padre”	8.6455

表 3 “XML not(retrieval) xtree” における拡張した検索結果を返す問合せ

問合せ	$score_{result}$
“XML not(retrieval) not(xtree) dataweb”	8.9009
“not(XML) retrieval not(xtree) db/ir”	8.7618
“not(XML) retrieval not(xtree) bm25f”	8.7618
“XML not(retrieval) not(xtree) padre”	8.6455
“not(XML) retrieval not(xtree) hyrex”	8.5591

(4) 式を用いることによって選出した上位 50 位までの拡張問合せ語を用いて問合せ語の拡張を行い、(10) 式を用いることによって検索結果を拡張することを試みた。拡張される検索結果を返す問合せの上位 5 位までについては表 2 のものが得られた。

“XML not(retrieval) xtree” という拡張された問合せに関しては “XTree for Declarative XML Querying” という検索結果が返され、“XML not(retrieval) dataweb” という拡張された問合せに関しては “A Designing Model of XML-Dataweb” という検索結果が返される。(10) 式の性質上、検索結果が少ないものが優先されているために、該当する拡張する検索結果としては 1 件のみのものやごくわずかなものがランキングの上位にきている。このことより、例えば利用者は xtree のような検索結果に関連のあるものを徐々に閲覧することが可能となり、検索結果について利用者はより深く理解することができる。

次に、“XML not(retrieval) xtree” という問合せで返ってきた検索結果の情報を閲覧した後に再度検索結果の拡張を行う状況を考える。“XML retrieval” という問合せで返ってきた情報に “XML not(retrieval) xtree” という問合せで返ってきた情報を追加して (4) 式を用いることによって、再度上位 50 位までの拡張問合せ語を選出して問合せ語の拡張を行い、(10) 式を用いることによって検索結果を拡張した。拡張される検索結果を返す問合せの上位 5 位までについては表 3 のものが得られた。

表 3 の結果は、表 2 から “XML not(retrieval) xtree” という問合せによって返ってきた検索結果を単に除去しただけのようにも思えるが、それはそもそも xtree に関する論文が少ないことに起因しているためである。利用者は提示された拡張された検索結果を徐々に閲覧していくことが可能となっており、増分的に検索結果を拡張していくことによって、利用者が興味のあるものを網羅的に閲覧することが可能となっている。

“XML retrieval” という問合せにおける拡張される語の候補上位 50 位までについて、(4) 式に関連のある (1) 式の E_S 、(2) 式の E_B の値に関する分布を調査したところ、図 7 が得られた。図 7 では横軸を E_B とし縦軸を E_S としている。ここで “hyrex” や “xtree” のような利用者が用いた “XML retrieval” という問合せ以外にはあまり関連がない問合せ語は、 E_B の値が大きく

(注2) : <http://dblp.uni-trier.de/xml/>

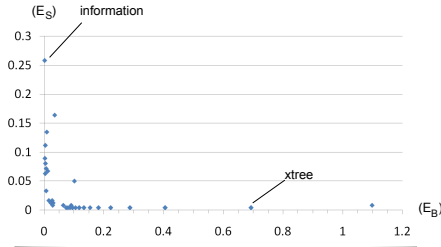


図 7 “XML retrieval” という問合せ語の拡張におけるスコア分布

表 4 “extraction feature” における拡張問合せ語の候補

拡張問合せ語	E_S	E_B	$score_{query}^{(4)}$
software	0.00088771	0.000046029	-0.00093373
management	0.00088771	0.000050333	-0.00093804
wireless	0.00088771	0.000064481	-0.00095219
service	0.00088771	0.000072440	-0.00096015
agent	0.00088771	0.000089043	-0.00097675

表 5 “extraction feature” における拡張した検索結果を返す問合せ

問合せ	$score_{result}$
“not(extraction) feature software”	18.108
“extraction not(feature) structured”	10.490
“not(extraction) feature modelling”	9.8006
“extraction not(feature) source”	8.4141
“extraction not(feature) user”	8.0912

E_S の値が小さい位置に分布されている。また，“content”や“information”のように利用者が用いた“XML retrieval”という問合せにも関連があるが，他の様々なキーワード集合にも関連のある問合せ語は， E_B の値が小さく E_S の値が大きい位置に分布されている。

さて，次に利用者が適切に問合せ語を作成できていない場合についても検討する。検索対象が曖昧な場合は，考えられ得る最適な問合せ語を作成することしかできず，もちろんその結果として適切に問合せ語を作成できる場合もあるが，できない場合もある。ここで，利用者が textmarker について調べたいが，その問合せ語を思い出せず，抽出に関するものだったことしか覚えていないとする。そこで，抽出の特徴について調べようとし，“extraction feature”という問合せで検索することを想定する。利用者が適切に問合せ語を作成できていないため，幅広い知識を必要としたために拡張問合せ語のスコア式として (7) 式を用いることを選択した場合，拡張問合せ語の候補上位 5 位までについては表 4 のものが得られた。

表 4 のように，拡張問合せ語の候補として，様々なキーワード集合と関連のある問合せ語が多く得られる。(7) 式を用いることによって選出した上位 50 位までの拡張問合せ語を用いて問合せ語の拡張を行い，(13) 式を用いて検索結果を拡張することを試みた。拡張される検索結果を返す問合せの上位 5 位までについては表 5 のものが得られた。

利用者は表 5 のような拡張された検索結果を閲覧することによって，textmarker では構造化されたデータも用いられていたことを思い出したとする。ここで，さらに“extraction not(feature) structured”という問合せによって返ってきた情

表 6 “extraction not(feature) structured” における拡張した検索結果を返す問合せ

問合せ	$score_{result}$
“extraction not(feature) not(structured) ausms”	8.9962
“extraction not(feature) not(structured) textmarker”	8.7935
“not(extraction) feature not(structured) irim”	8.7513
“not(extraction) feature not(structured) foley”	8.2609
“extraction not(feature) not(structured) trecvid2008”	8.1915

報を基に増分的に検索結果を拡張するが，その返ってきた情報と関連が強く，なおかつ拡張量を小さくしたい場合と想定する。拡張問合せ語のスコア式として (4) 式を用いることを選択した場合，拡張される検索結果を返す問合せ語の上位 5 位までについては表 6 のものが得られた。

拡張された検索結果を閲覧することによって，利用者は求めていた textmarker という情報が得られる。

6. ま と め

本稿では，利用者によって用いられた問合せ語を基に，その問合せ語および検索結果を拡張する手法を提案した。問合せ語の拡張には情報量を用いることによって，利用者が適切に問合せ語を作成できている場合と作成できていない場合の両方に効果的な拡張問合せ語を抽出している。そして，選出された拡張問合せ語を用いて問合せ語の拡張を行い，拡張する検索結果を，拡張量や idf 値を考慮することによってランキングを行った。また，実験において事例を挙げて有効性を検討した。

今後の課題として，構造化データおよび構造化問合せへ本手法を拡張することが考えられる。構造化データを扱う際には，例えば DBLP のような文献データベースにおいて，著者やタイトルというような属性情報を用いることで，より高度な検索結果の拡張手法を適用することが考えられる。また，DBLP 以外のデータセットを用いた検証を行い，さらに，拡張した検索結果を利用者が効果的に閲覧可能となっているのか，つまり本手法の定量的な評価についても検討していかなければならない。

文 献

- [1] 堀幸雄, 今井慈朗, 中山亮. ユーザの web 閲覧履歴を用いた検索支援システム. 情報知識学会誌, Vol. 17, pp. 95–100, 2007.
- [2] 岡部正幸, 山田誠二. トランスダクティブ学習による最小文書判定からのクエリ拡張. 人工知能学会論文誌, Vol. 21, No. 4, pp. 398–405, 2006.
- [3] 佐藤省吾, 田中二郎. メタデータに基づく検索/閲覧インタフェース. Master’s thesis, 筑波大学大学院理工学研究科, 2006.
- [4] 増井俊之, 塚田浩二, 高林哲. 近傍関係にもとづく情報検索システム. 第 11 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS2003), pp. 79–86, 2003.
- [5] Ziyang Liu, Peng Sun, and Yi Chen. Structured search result differentiation. *Proceedings of the Very Large Data Base (PVLDB)*, pp. 313–324, 2009.
- [6] Quoc Trung Tran Chee-Yong Chan and Srinivasan Parthasarathy. Query by output. *Special Interest Group on Management of Data (SIGMOD)*, pp. 535–548, 2009.
- [7] Joaquin Perez-Iglesias, Jose R. Perez-Aguera, Victor Fresno, and Yuval Z. Feinstein. Integrating the probabilistic model bm25/bm25f into lucene. *Clinical Orthopaedics and Related Research (CoRR) abs/0911.5046*, 2009.