キーワードを利用した XML 文書検索のための検索結果粒度決定法

Determining Fine-grained Results for Keyword-based XML Document Retrieval

波多野 賢治♡ 絹谷 弘子◇
吉川 正俊♠ 植村 俊亮♡

Kenji HATANO Hiroko KINUTANI
Masatoshi YOSHIKAWA
Shunsuke UEMURA

本論文では，キーワードを利用した XML 文書検索システムの実現のために，あらかじめ XML 文書を分割し検索結果候補とする際の検索結果候補の粒度決定法について述べる．提案する部分文書の粒度決定法によって，検索結果候補の数を削減することが可能であるため，XML 文書検索の高速化およびその高精度化を図ることが可能となる．

This paper proposes a method for determining fine-grained results for keyword-based XML document retrieval. Determining a granule of retrieval results, the number of targeted retrieval results of XML documents will be reduced, so that retrieval time will be reduced and overall performance of XML document retrieval system will be boosted.

1. はじめに

XML (Extensible Markup Language) [3] が，情報化社会に与えた影響は非常に大きく，世間では WWW (World Wide Web) に次ぐ大きな提案であったとまで言われており，非常に多くのアプリケーションで XML が用いられるようになってきている．このような背景から，計算機上に存在するあらゆるデータが，近い将来，XML 形式で記述されるだろうと考えられ，WWW の発展に伴って WWW 検索システム (Web 検索エンジン) が開発されたように，XML 文書検索システムへの期待は大きくなると予想される．

XML 文書を検索するための手法の標準は XML 問合せ言語 [2] であり，これらは市販の XML 対応を謳ったデータベースの検索機能に盛り込まれたり，W3C (World Wide Web Consortium) からワーキングドラフトが公開されたり [1] と，盛んに研究が行われている．しかし，これら XML 問合せ言語は，データベース問合せ言語の SQL と同様，問合せを行うための専門的知識や，利用者があらかじめ検索したい XML 文書の文書構造を把握し検索の際にそれらを指定する必要があるため，利用者の利便性を考えると Web 検索エンジンのように使いやすいものとはいえないのが現状である．

♡ 正会員　奈良先端科学技術大学院大学 情報科学研究科
{hatano,uemura}@is.aist-nara.ac.jp
◇ 正会員　科学技術振興事業団 戦略的創造研究推進事業
kinutani@dblab.is.ocha.ac.jp
♠ 理事　名古屋大学 情報連携基盤センター
yosikawa@itc.nagoya-u.ac.jp

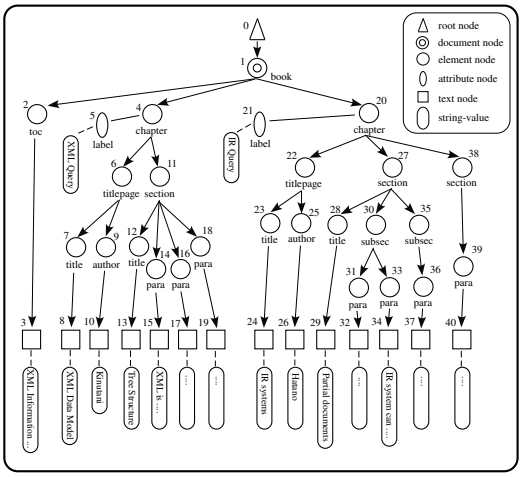

図 1: XML 文書の木構造表現

このような利用者に対する利便性に関する問題点を克服するために，我々はこれまで，問合せキーワードを入力するだけで利用者が求めている XML 文書を検索できるシステムを開発してきた [6, 8]．開発したシステムでは，Web 検索エンジンのように，利用者は検索キーワードを入力するだけで求めている情報，すなわち，問合せに相応しい XML 文書中の一部分を検索することができ，さらにそれらは問合せ内容に対する相応しさを基にランキングされている．しかし，キーワード入力による問合せの実現のために，検索対象となる XML 文書をその文書構造を利用してあらかじめ XML 部分文書に分割する必要があるため，検索対象 XML 部分文書数が膨大となり，それとともに検索に要する時間も膨大となるという問題点を持っていた．

こうした問題点を解決するために，本論文では検索対象となる XML 部分文書の粒度をあらかじめ決定することで，検索対象 XML 部分文書数を削減する手法を提案する．我々は，検索対象となる XML 部分文書には 2 種類，すなわち利用者にとって有益な内容を含んでいる部分文書 (以下 CPD (Coherent Partial Document) と表記する) と不要な部分文書[1]があると考えており，不要な部分文書を検索対象から除外することで，検索対象 XML 部分文書の数を削減し検索の高速化を図る．

2. CPD (Coherent Partial Document)

2.1 XML 部分文書

XPath データモデル [5] において[2]，XML 文書は階層構造をもった木構造で表現され，それぞれの節点は document order を利用して ID が振られている．木構造の葉 (leaf node) は，図 1 のように text node もしくは attribute node であり，根ノードの子は document node と呼ばれている．また，document node と leaf node 間にある中間ノードは element node と呼ばれている．この XPath データモデルに基づいた XML 文書のための検索モデルは，これまでに 2 種類提案されているが，本論文における検索モデルは proximal node モデル [10] に類似しており，その検索モデルを利用して，XML 部分文書の定義を以下のように定めている．

定義 1 (XML 部分文書) *XML* 文書中に出現するすべての要素について，開始タグと終了タグで囲まれた部分，すなわち，*document node* または *element node* を根とする木全体を *XML* 部分文書と

[1] このような部分文書のことを，文献 [7] では stop-contexts と呼んでいる．この文献においても，検索システムの scalability の確保には stop-contexts の除去が必要であると述べられている．

[2] XPath データモデルで扱われている 7 種類のノードのうち，本論文では document node, text node, attribute node, element node に限定している．

呼ぶ．本稿ではこのような *XML* 部分文書を，その根につけられている *ID n* を利用して *XML* 部分文書 *#n* と呼ぶ．

2.2 CPD

利用者にとって意味のある XML 部分文書，すなわち CPD とは，文書構造および文書内容について意味的にまとまりのある部分文書であり，従来の情報検索技術で検索される単に検索キーワードを含んだ XML 部分文書とは異なる．精度のよい，しかも利便性の高い検索システムを構築するためには，CPD を検索対象とすべきである．

例えば，入力キーワードとして `Hatano` を従来型のパッセージ検索システムに与えた場合，図 1 の XML 文書からその検索結果として，XML 部分文書 `<author>Hatano</author>` が返される．この XML 部分文書は，利用者が必要としているキーワードを含んでいるが，`Hatano` が何の `author` であるか示されていないため，利用者にとっては情報量が不足しており，検索結果としては不適切である．また，従来型の文書検索システムのように，図 1 が示す XML 文書全体が先の問合せの検索結果として返されても，利用者にとって問合せの解として不必要な 1 番目の `chapter` の情報まで含まれ，情報過多な検索結果であるため不適切だと考えられる．

図 1 が示す XML 文書中に含まれる部分文書のうち，先に例として挙げた検索要求に最も相応しいと思われる部分文書，すなわち意味のある XML 部分文書は，要素 ID #20 を root node とする XML 部分文書 #20 である．なぜなら，この XML 文書には 2 つの `chapter` が存在し，`Hatano` は 2 番目の `chapter` の `author` だからである．利用者が情報検索を行う場合は，入力キーワードを含んでいる最小の部分文書ではなく，XML 部分文書 #20 のような意味のある XML 部分文書群を検索対象とすべきであり，そのことが検索精度を向上させ，また検索システムの利便性の向上にも結びつく．本論文では，このような XML 部分文書のことを CPD と呼ぶが，XML 文書中から分割されるすべての部分文書が CPD に該当するわけではない．そのため，従来の文書検索と同様に利用者が文書構造を意識せずに検索要求として検索キーワードを与えるだけで，これら CPD を検索結果として得るためには，検索システムに CPD を決定する仕組みが必要となる．

CPD を決定する仕組みとして，我々は XML 文書構造を利用して CPD を決定する手法を提案した [8]．この手法では，CPD は文書の論理構造によって決定されるものであり，CPD を表現する XML 部分文書の root node は，その兄弟ノードに同名の要素名をもつことが多いという事実を利用している．しかし XML 文書内には，文書の論理構造だけではなく語の強調やリンクのアンカーなどに用いられる要素も多数存在するため，CPD として抽出されるべき XML 部分文書が抽出されないなどの問題点があり，あらゆる XML 文書に適用することができないことが判明した．

本論文では，検索対象となる XML 文書の文書構造の性質だけを利用するのではなく，XML 文書の持つ統計量，例えば XML 部分文書自身に含まれる単語数や異なり語数などを利用して明らかに CPD とはなりえない XML 部分文書を除去し，残った XML 部分文書を CPD であるとする新しい検索対象 XML 部分文書の粒度決定法を提案する．

3. 統計量を利用した分析

3.1 プロトタイプシステム

1 章でも述べたように，我々は利用者に対する利便性を考慮し，問合せキーワードを入力するだけで利用者が XML 部分文書を検索できるシステムを開発してきた [6, 8]．図 2 に開発中の XML 部分文書検索システムの概略図を示している．図に示したように，我々の提案システムでは XML 文書を XML パーザー Xerces[3] を用いて

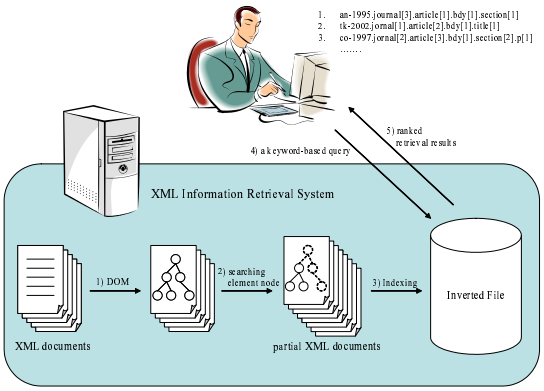

図 2: XML 文書検索システムの概略図

DOM 木を構築する部分，構築された DOM 木から element node を探索する部分，探索された element node を根とする XML 部分文書の索引ファイルを構築する部分，そして利用者の問合せに対し各 XML 部分文書と問合せとの類似度をベクトル空間モデルにしたがって計算し，それらを基にランキング付きの検索結果を提示する部分の 4 つから構成されている．

我々が文献 [6] で提案した XML 文書検索システムでは，XML 文書木中の element node を根とする全ての XML 部分文書を検索対象としていたが，本論文で提案する手法は，XML 部分文書の統計量を利用して有益な内容を含んでいる XML 部分文書だけを検索対象にするように改善する．このような有益な内容を含んでいる XML 部分文書を，本論文では CPD として扱う．

分析に使用した XML 文書は，IEEE Computer Society から 1995 ～ 2002 年に発行された雑誌および論文誌に含まれている記事および論文であり，含まれている論文数は 12,107 文書である．この XML 文書群は，2002 年 4 月に発足した INEX Project[4] が INEX test collection として使用しており，すべての記事，論文が DTD (DTD の制定は，INEX Project ではない) に基づいて論理的な 1 つの XML 文書として表現されている．DTD 中で定義されている文書要素は 192 種類であり，その XML 文書サイズは 496 MBytes にのぼる．

3.2 統計量の分析

本論文では，XML 文書から抽出することが可能な統計量として XML 文書が持つ文書構造にしたがって分割した XML 部分文書に含まれる単語数，異なり語数，そして単語数と異なり語数から計算される異なり語率を利用した．この 3 種類とした理由には，XML 部分文書は単語で構成されており，また，その内容は文章，数式，固有名詞を含む単語，数値など多彩であるため，XML 部分文書に含まれる文数など単語に関係ない統計量を利用することが難しいからである．なお，異なり語率は以下のように定義する．

定義 2 (異なり語率) *XML* 部分文書中に出現する単語数を n^w，異なり語数を n^k とすると，異なり語率 R は以下のように表現される．

$$R = \frac{n^k}{n^w} \tag{1}$$

異なり語率を定義する理由は，XML 部分文書に含まれている単語数はさまざまであるため，検索を行う際に XML 部分文書と問合せとの類似度をベクトル空間モデルで評価するのに適しているかどうかを判定するためである．一般にベクトル空間モデルで評価可能な文書には，同じ単語が何度も含まれており，異なり語率は 100% とはならない．その一方，カタログのデータ一つ一つを表す文書には，同じ単語が複数出現することはほとんど考えられ

[3] `http://xml.apache.org/xerces-j/index.html`

[4] Initiative for the Evaluation of XML Retrieval (INEX): `http://qmir.dcs.qmul.ac.uk/INEX/`.

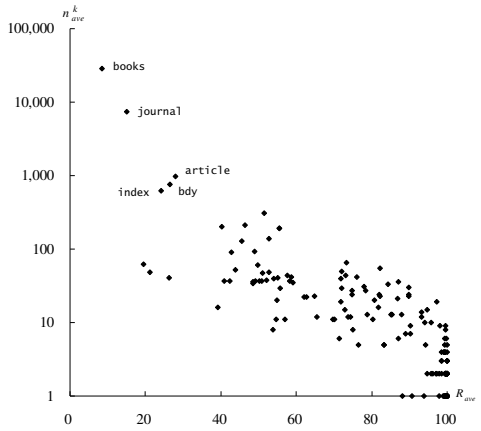

図 3: 平均異なり率 R_{ave} と平均異なり語数 n^k_{ave} の関係

ず，異なり語率は 100% に近いといえる．すなわち，異なり語率が 100% に近い XML 部分文書はデータ指向が強いと考えられるため，それらは提案システムでは CPD として抽出され検索対象となるべきではないと考えられる．したがって，以下のような手順で統計量から CPD を決定していく．

1. 図 2 に示したように，INEX test collection を表現する XML 文書を Apache Xerces を利用して DOM 木に展開し，さらに，その element node を探索しておく．探索された element node には document order にしたがって ID が付けられる．
2. 抽出した element node を根とする XML 部分文書を XML 文書から切り出す．2.1 節で述べた XML 部分文書の定義から，XML 文書から XML 文書中の element node の数と同数の XML 部分文書が抽出されることになる．
3. 各 XML 部分文書に含まれる単語数 n^w，異なり語数 n^k，そしてそれらの比を表す異なり語率 R を利用して，CPD として相応しい XML 部分文書を決定していく．具体的には n^w，n^k，そして R においてそれぞれある閾値を設定し，その閾値を利用して XML 部分文書が CPD として相応しいかどうかを決定する．この際，ストップワード処理や接辞処理などの前処理はあらかじめ行った上でそれぞれの統計量を利用する．こうして決定された CPD の文書数 N が，検索対象 XML 部分文書数となるので，N を利用することで XML 検索システムの高速化の指標となる．

3.3 統計量の分析結果および考察

統計量の解析結果を散布図にまとめたものを図 3 に示す．図中の平均異なり語率 R_{ave} は，XML 部分文書の root node 名が同じである XML 部分文書 d_i が持つ単語数を n^w_i，異なり語数を n^k_i としたとき，

$$R_{ave} = \frac{\sum_i n^k_i}{\sum_i n^w_i} \quad (2)$$

で計算される値を表している．

図 3 が示すように，INEX test collection 全体を表現する XML 文書の根に近い要素 (例えば，books，journal，article など) を root node とする XML 部分文書には，多くの単語，異なり語が含まれており，それらの多くはその平均異なり語率 R_{ave} が小さい．これに対して，平均異なり語率 R_{ave} の値が比較的高い XML 部分文書に含まれている平均異なり語数 n^k_{ave} は 100 語未満であり，平均異なり語率 R_{ave} の値が大きい XML 部分文書ほど XML 部分文書のサイズは小さいことがわかる．

また，部分文書の root node ごとに集計した平均異なり語率 R_{ave} の値によって 11 のグループに分類して，統計量の分析を詳細に行ってみると (図 4 参照)，全体の約 3 割にあたる 62 種類の

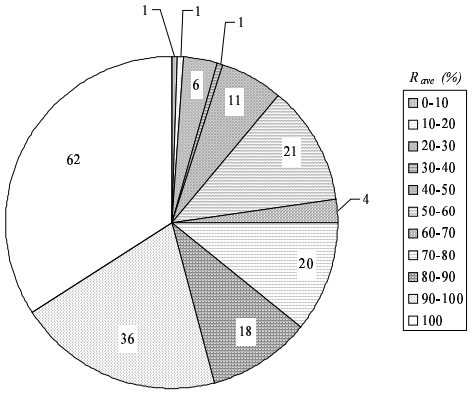

図 4: 平均異なり率 R_{ave} による XML 部分文書の分類

XML 部分文書の平均異なり語率 R_{ave} が 100%，また全体の約 2 割にあたる 36 種類の XML 部分文書の平均異なり語率が 90% 以上 100% 未満であった．これらの XML 部分文書の多くは，INEX test collection の XML 文書木において葉の部分にあたり，また，それらに含まれている単語数，異なり語数は非常に少ない．

これらの結果は 3.2 節を踏まえると，平均異なり語率 R_{ave} が高い XML 部分文書はデータ指向が強い部分文書であり，本論文で実装しているキーワードを利用したベクトル空間モデルに基づく検索システムの検索結果として用いられるべきではない部分文書であるといえる．すなわち，XML 部分文書の持つ平均異なり語率 R_{ave} を変化させることで検索対象 XML 部分文書数の調整が可能であるため，検索システムの課題であった高速検索の実現を，異なり語率 R_{ave} を利用して実現することが可能であることが分かった．

以上の点から，CPD は XML 部分文書の平均異なり語数 n^k_{ave} と平均異なり語率 R_{ave} によってある程度絞り込むことが可能であり，CPD の条件として以下の点を考慮することが有効であると思われる．

- 平均異なり語率 R_{ave} が 90% 以下の XML 部分文書のほとんどは，その部分文書中に 1,000 語以下の異なり語を含んでいる．一般的にサイズの大きな文書全体を検索結果とすることは，利用者が検索結果を閲覧する際に検索要求に合致する部分を，検索結果から再度見つける必要があり非常に不便である．したがって，CPD に相応しい XML 部分文書の条件として，平均異なり語率を導入することも有益であり，INEX test collection の場合は，1,000 語以下の XML 部分文書を検索対象とすべきである．
- 文献 [8] で提案した文書構造を利用した CPD の抽出手法では，CPD を表現する XML 部分文書の root node に対し同名の兄弟ノードを持つことが多いという事実を利用していた．この CPD の持つ特長は本論文においても有効な決定手法であるため，XML 部分文書の出現数 N の値が大きく，またその平均異なり語率 R_{ave} が小さな部分文書は CPD として相応しいと考えられる．
- 3.2 節で述べたように，平均異なり語率 R_{ave} が 100% の XML 部分文書はデータ指向が強い部分文書であるため，ベクトル空間モデルによって正確に検索することができない．したがって，そういった XML 部分文書は CPD として検索対象となるべきではないと考えられる．平均異なり語率 R_{ave} の閾値を決定するためには，INEX test collection の query/answer セットが必要となるが，例えば，平均異なり語率 R_{ave} が 90% 未満の XML 部分文書を CPD とすれば，CPD として定義される XML 部分文書数は INEX test collection を表現する XML 文書木から抽出される XML 部分文書数の約 3 割に減少し (図 5 参照)，XML 文書検索システムの高速検索が実現可能となる．

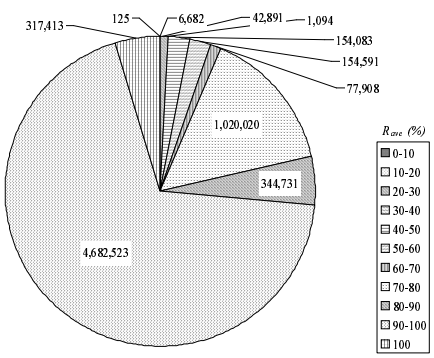

図 5: 平均異なり率 R_{ave} による XML 部分文書数 N

4. 関連研究

文書検索の研究分野において，検索要求に対してそれに類似した文書の一部分だけを検索するという研究テーマは，パッセージ検索 [11] が提案されてから非常に注目されている．これらの研究の主眼は，文書の一部分を検索することに置かれているが，見方を変えれば検索対象の文書の粒度 (単位) をどのように決定するかについて提案しているとも言え，単に検索精度を向上させるためだけではなく，検索システムのパフォーマンスの確保などにも利用されている．

近年，特にこれらの研究テーマが盛んに行われているのは，Web 文書検索の分野であり，文献 [9] や [12] では Web 文書間に張られているリンクを利用して文書間の関連度を計算し，それを基に Web 文書検索における検索対象文書粒度 (Information Unit) を決定しようとしている．また，半構造データにおいても同様の研究が始められており [4]，XML 文書検索の分野においても検索精度の向上だけではなく，検索システムの scalability の確保やパフォーマンスの向上などさまざまな効果が期待される研究テーマである．

5. おわりに

本論文では，問合せキーワードを利用した XML 文書検索システムを構築する際に生じる，検索対象 XML 部分文書数が膨大となることによる検索コストが増加するという問題に対して，XML 部分文書から抽出される単語数などの統計量を利用した検索対象 XML 部分文書の粒度決定法を提案した．また提案した手法を利用すれば，抽出可能な XML 部分文書の 3 割程度に文書数を抑えることができ，より高速な検索が実現可能であることが確認できた．本論文で提案した CPD の概念は，検索対象となる XML 文書が大きくなればなるほど検索システムの高速化を図るために必要であり，さらに検索精度を高めるために有効な手法であると考えている．

今後の課題としては，本論文で判明した CPD の条件を，さらに INEX test collection の query/answer セットを利用してより詳細に決定し，それを適用することによる，XML 文書検索システムの検索時間短縮の効果および検索精度の向上の確認，および，CPD の決定条件に利用した統計量について，計量情報学における理論的な裏づけをとることが挙げられる．

[謝辞]

本研究の一部は，文部科学省科学研究費若手研究 (B) (課題番号: 14780325) および科学技術振興事業団 (JST) の戦略的基礎研究推進事業 (CREST) 「高度メディア社会の生活情報技術」プログラム機構の支援によるものである．

[文献]

[1] S. Boag, D. Chamberlin, M.F. Fernandez, D. Florescu, J. Robie, and J. Siméon. XQuery: A Query Language for XML. http://www.w3.org/TR/xquery, Nov. 2002. W3C Working Draft 15 November 2002.

[2] A. Bonifati and S. Ceri. Comparative Analysis of Five XML Query Languages. *ACM SIGMOD Record*, Vol. 29, No. 1, pp. 68–79, Mar. 2000.

[3] T. Bray, J. Paoli, C.M. Sperberg-McQueen, and E. Maler. Extensible Markup Language (XML) 1.0 (Second Edition). http://www.w3.org/TR/REC-xml, Oct. 2000. W3C Recommendation 6 October 2000.

[4] S. Chakrabarti. Text Search for Fine-grained Semi-structured Data. In *Tutorial Notes of the 28th International Conference on Very Large Data Bases*, pp. 115–135, Aug. 2002.

[5] J. Clark and S. DeRose. XML Path Language (XPath) Version 1.0. http://www.w3.org/TR/xpath, Nov. 1999. W3C Recommendation 16 November 1999.

[6] K. Hatano, H. Kinutani, M. Yoshikawa, and S. Uemura. Information Retrieval System for XML Documents. In *Proc. of the 13th International Conference on Database and Expert Systems Applications*, Vol. 2453 of *LNCS*, pp. 758–767. Springer-Verlag, Sep. 2002.

[7] G. Kazai and T. Rölleke. A Scalable Architecture for XML Retrieval. In *Proc. of the First Workshop of the Initiative for the Evaluation of XML Retrieval*. ERCIM, Mar. 2003. (to appear).

[8] 絹谷弘子, 波多野賢治, 吉川正俊, 植村俊亮. XML 文書の文書構造と内容を用いた部分文書の抽出手法. 情報処理学会論文誌: データベース, Vol. 43, No. SIG2(TOD13), pp. 80–93, Mar. 2002.

[9] W.-S. Li, K.S. Candan, Q. Vu, and D. Agrawal. Retrieving and Organizing Web Pages by "Information Unit". In *Proc. of the 10th International World Wide Web Conference*, pp. 230–244, May 2001.

[10] G. Navarro and R. Baeza-Yates. Proximal Nodes: A Model to Query Document Databases by Content and Structure. *ACM Transactions on Information Systems*, Vol. 15, No. 4, pp. 400–435, 1997.

[11] G. Salton, J. Allan, and C. Buckley. Approaches to Passage Retrieval in Full Text Information Systems. In *Proc. of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–58, June/July 1993.

[12] K. Tajima, K. Hatano, T. Matsukura, R. Sano, and K. Tanaka. Discovery and Retrieval of Logical Information Units in Web. In *Proc. of the 1999 ACM Digital Library Workshop on Organizing Web Space*, pp. 13–23, Aug. 1999.

波多野 賢治 Kenji HATANO
奈良先端科学技術大学院大学情報科学研究科助手．情報検索システム，データベースシステムの研究に従事．情報処理学会，電子情報通信学会，日本データベース学会正会員．

絹谷 弘子 Hiroko KINUTANI
科学技術振興事業団戦略的創造研究推進事業研究員．情報検索システム，データベースシステムの研究に従事．情報処理学会，日本データベース学会正会員．

吉川 正俊 Masatoshi YOSHIKAWA
名古屋大学情報連携基盤センター教授．データベースシステムの研究に従事．情報処理学会，電子情報通信学会正会員．日本データベース学会理事．

植村 俊亮 Shunsuke UEMURA
奈良先端科学技術大学院大学情報科学研究科教授．データベースシステムの研究に従事．情報処理学会，電子情報通信学会フェロー．IEEE Fellow．日本データベース学会正会員．著書に「データベースシステムの基礎」(オーム社) など．