質問の階層的構造化を用いたWeb検索手法の提案

Web Search Using Hierarchical Structuring of Queries

小山 聡* 田中 克己*

Satoshi OYAMA Katsumi TANAKA

ユーザが検索エンジンに複数のキーワードを入力した場合，これらのキーワードには，対象となる話題を表現する際の異なった役割が想定されている場合が多い．例えばあるキーワードは主題を表し，別のキーワードは主題に関する内容を記述するといったような場合である．同じキーワードからなる検索式であっても，このようなキーワードの役割を区別することで，検索精度を向上できる可能性が考えられる．本論文では，既存の検索エンジンの検索機能を利用し，ページのタイトルに含まれるキーワードと本文中に現れるキーワードを区別した 2 種類の質問の検索結果を比較する予備実験を行った．結果は，二つの質問で検索されるページの内容に違いが現れるというものであった．そこで，我々は予備実験の設定を一般化し，質問の階層的構造化を用いた Web 検索方式を提案する．さらに，提案手法の応用例として，質問の修正に用いる方式と，検索結果の比較呈示に用いる方式の検討を行う．

When a user inputs multiple keywords to a search engine, he usually assumes these keywords to have some different roles in describing the topic of interest. For example, the user may use some keyword as the subject and use other keywords for describing a certain aspect of the subject. By treating keywords differently according to their roles in the query, there is a possibility of improving the precision of search results. We conduct preliminary experiments by submitting two different types of queries to an existing search engine: queries that require the one keyword in the title and the other one in the body, and otherwise. The results are encouraging because the results of these two types of queries have difference in their contents. Therefore, we generalize the setting of the experiments and propose a new method of web search using hierarchical structuring of queries. We also present the applications of our method to query modification and comparative presentation of search results.

1. はじめに

Webによって提供される情報の量や種類はますます増加し，我々は日常生活や仕事で必要な情報の多くを，Webから得ることが可能になった．大量のWebページの中から，自分の必要な情報が掲載されているページをみつけるための最も一般的な手法は，検索エンジンを用いることである．しかし，Webが様々な話題を含むため，ユーザの入力した質問に対して，しばしばユーザの本来の要求とは無関係の多くのページまでが検索されてしまうという問題が生じる．

必要とするページだけが検索されるような適切な質問を記述することは，ほとんどのユーザにとって困難な作業である．そこで，ユーザの質問作成を支援するさまざまな手法が試みられている．例えば，幾つかの検索エンジンでは，ユーザの入力したキーワードに関連のあるキーワードを複数呈示し，ユーザが検索の絞込みに使える機能を提供している．

また，“検索隠し味”法[1]では，ドメインに固有のキーワードのブール式をユーザの入力キーワードにAND演算子で付加し，ブール式の入力を受け付ける汎用のWeb検索エンジンに投入することで，Webページの中から特定のドメインに関するページだけを検索できるようにしている．

これらの方法に共通するのは，新しいキーワードを追加することで，検索性能を向上させようという考え方である．ここでは，ユーザの入力したキーワードと追加されたキーワードはAND演算子で結ばれ，その関係は対等であった．また，ユーザが複数キーワードを入力した場合も，それらのキーワード間の関係は対等に扱われていた．

しかし，通常は，これらのキーワードの間には，主題となるキーワードと，主題の特定の側面を記述するために付加されたキーワードといったように，非対称な関係が成り立つ場合が多い．すなわち，同じキーワードから成る質問であっても，全く異なる意味的な構造を持ちうることになる．また，Webページの中においても，ページのタイトルに現れるキーワードと，本文中に現れるキーワードとでは，そのページにおける役割は異なっていると考えられる．

そこで，本論文では，ユーザの質問の中の主題的なキーワードと，付加的なキーワードを区別し，質問を構造化し，Webページの構造とのマッチングを行うことで，同じキーワードからなる質問を用いて，検索精度を向上させることを試みる．

オプション	機能
intitle:	タイトルに含まれる語を検索
intext:	本文に含まれる語を検索

表1 Google における検索オプション
Table 1 Search Options of Google

2. 予備実験

ここでは，既存の検索エンジンの機能を利用した予備実験により，質問中のキーワード間に構造を与えることで，検索結果にどのような影響が現れるかを評価・考察する．

HTMLで記述されたWebページは，単なるテキストと異なり，構造を持っている．そこで，この構造を検索の際に利用することを考える．例えばGoogle[1]では，表1のような検索オプションを提供している．

そこで，ユーザからの質問が二つのキーワードA,Bからなっている場合を想定し，キーワードAがタイトルに，キーワードBが本文に含まれている質問,およびキーワードBがタイトルに，キーワードAが本文に含まれている質問の二種類を用意し，Googleに投入することによって検索結果を比較した．

* 正会員 京都大学大学院情報学研究科社会情報学専攻
{oyama,ktanaka}@i.kyoto-u.ac.jp

1 http://www.google.com

intitle:プラスチック intext:環境保護	intitle:環境保護 intext:プラスチック
普及に向けて開発進む生分解性プラスチック	生産活動での環境保護
プラスチック製振込カード／TTT Card	環境保護の取り組み
伊藤忠プラスチック・システム：オランダ	環境保護の取り組み
伊藤忠プラスチック・システム： …	NTT 環境保護活動報告 2000 第 4 章 第 2 節 第 3 項 超 …
小幡谷･伊藤グループ（研究紹介/生分解性 …	環境保護活動報告 2001 － リサイクル推進
浦谷商事株式会社：製品紹介：プラスチック …	環境保護
双分解プラスチック	ガスタービン燃焼器からゴミ焼却まで －環境 …
双分解プラスチック	環境保護用語集
世界初の廃プラスチック高炉一貫リサイクル …	生産活動での環境保護
「廃プラスチック高炉原料化設備」 …	生産活動での環境保護

表 2 異なった構造を持つ質問間の検索結果の比較

Table 2 Comparison of search results of queries with different structures

例えば，“プラスチック”と“環境保護”という2つのキーワードを用いる場合，“intitle:プラスチック intext:プ環境保護”および“intitle:環境保護 intext:プラスチック”という2種類の質問を投入した．この場合，前者に対しては54件，後者に対しては60件のページがヒットした．また，これらの検索結果の中のURLを比較したところ，二つの検索結果両方に含まれるURLは1件もなかった．

また，“大学”と“金融工学”という組み合わせでは，“intitle:大学 intext:金融工学”に対して112件，“intitle:金融工学 intext:大学”に対して86件の検索結果があったが，両方の検索結果に共通するURLは3件であった．これらは，質問の構造によって，検索されるURLの集合が大きく異なる場合があることを示している．

また，単にURLの重なりだけでなく，検索されたページの内容自体も，異なった傾向がみられる場合がある．例えば，表2のように，“プラスチック”と“環境保護”の場合，“プラスチック”をタイトルに指定した場合，分解されやすいプラスチックや，環境保護を意識したプラスチック製品に関するページが多く現れている．一方，“環境保護”をタイトルに指定した質問では，環境問題のページの中でプラスチックの問題を述べているページが多く現れており，二つの質問で検索されるページの内容が異なっていることが分かる．また，“大学”と“金融工学”の例では，“大学”をタイトルに指定した質問の結果においては，金融工学の講座や授業を持つ大学が多く現れ，“金融工学”を指定した質問の結果においては，学会や講演，解説などの金融工学一般の話題を持つページが多く現れている．

この実験は，あくまで少数の例に対し，1つの検索エンジンで行った予備実験であるが，質問の構造によって検索結果のURLや内容が大きく異なる場合があることを示している．そのような場合には，ユーザの入力した質問に適切な構造を与えてやることで，検索結果を向上できる可能性があると考えられる．

3. 質問の階層的構造化を用いた Web 検索

3.1 ページの階層構造のモデル化

前節の結果により，タイトルに含まれるキーワードと本文に含まれるキーワードを区別して取り扱うことで，検索結果を変化させうることが明らかになった．本節では，予備実験の問題設定を一般化し，質問の階層的構造化を用いた Web 検索方式の提案を行う．

予備実験では，Web ページのタイトルと本文という一段階の構造しか考慮していない．これは，既存の検索エンジンが，タイトルもしくは本文に検索範囲を限定した検索機能しか提供していないからである．しかし，実際の Web ページは，ページの中が幾つかの部分に分かれており，それぞれの部分がまたサブタイトルや本文で構成されるといった，階層構造を持っている場合が多い．例えば，キーワード“プラスチック”がサブタイトルに含まれており，その下のテキストに“環境”が含まれているような場合，ページのこの部分は，タイトルに“プラスチック”，本文に“環境”が含まれているページと同じ構造を持っているとみなすことができる．そこで，このようなパターンも検索できる方式を考えたい．

そのために，まず，Web ページの階層構造を図 1 のような木で表すことを考える．すべての節点はキーワードの組からなり，根節点はページのタイトル，葉節点は文章から取られたキーワードである．中間節点は，ページ内のサブタイトルからとられたキーワードで記述する．ただし，ある節点で現れたキーワードはその子孫の節点には現れないように予め取り除いておく．

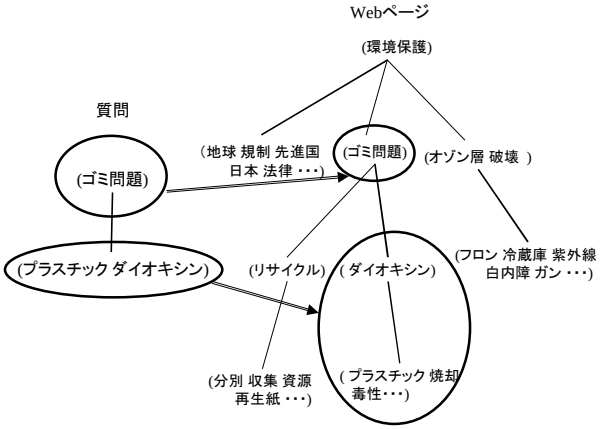

図 1 質問と Web ページのマッチング

Fig.1 Matching a Query with a Web Page

3.2 包含関係を用いた質問とページとの合致判定

検索を行うためには，どのような場合に質問とページが合致するのかを定義する必要がある．ここでは，検索式も Web ページと同様のキーワードの階層で表されているとし，質問とページの合致条件を考える．

ここで，各節点がキーワードの集合であるような二つの木が与えられたときに，片方をもう一方が含むという包含関係を関係のように定義する．

まず，ある一つの節点（キーワード集合）v と，節点の集合（キーワード集合族）$S = \{u_1, u_2, \ldots, u_n\}$を考える．このとき，$v$ の中の全てのキーワード $k \in v$ が $k \in u_1 \cup u_2 \cup \Lambda \cup u_n$ を満たすとき，v は S で被覆されるとい

う．また，同じ木の節点 v_i と v_j について，v_i が v_j の先祖であるとき，$v_i \pi v_j$ と記述する．

このとき，以下が成立する場合に木 A は木 B に含まれると定義する．

- 木 A の全ての節点 $v_1,v_2,\ldots,v_n$ に対して，v_i がそれぞれ S_i に被覆されるような $S_1, S_2,\ldots, S_n$ が木 B 上に存在する．
- 木 A において二つの節点の間に $v_i \pi v_j$ の関係が成り立つとする．上の条件で，v_i が S_i に，v_j が S_j に被覆される時，木 B において，任意の $u' \in S_i$ と $u'' \in S_j$ の間に $u' \pi u''$ が成り立つ．

上に定義した木の間の包含関係を用いれば，図 1 の例のように，質問を表現した木がページを表現した木に含まれるような場合に，検索結果とすることができる．この場合，質問においてキーワード k_1 の含まれる節点よりも先祖の節点に含まれていたキーワード k_2 は，検索結果のページの中でも先祖の節点に含まれる．

ここでの包含関係の定義は，木 A が木 B の部分木であるという意味での通常の包含関係よりは，緩い条件になっている．これは，検索の際には，キーワードの間の先祖関係が保存されていればよく，厳密に構造が一致することを要求すれば，検索できるページ数が少なくなってしまうと考えるからである．

4. 質問の修正への応用

上の例のような検索を行うためには，質問が構造化されている必要がある．最も直接的な方法は，ユーザにキーワード間の関係を明示的に指定してもらうことである．例えば，AND, OR, NOT といった通常の検索エンジンで用いることができる演算子に加えて，新しい演算子 AS といったものを導入し，“プラスチック AS 環境問題”といった質問を入力してもらえば．システムは"プラスチック"が主題となるキーワードで，"環境問題"が主題に関する内容を記述するキーワードであると解釈することができる．しかし，単純なブール式でさえ多くのユーザは使いこなすことが困難であるといわれており[2], ユーザに新しい演算子の利用を求めることは難しいと考えられる．3 語以上からなる質問をユーザが構造化することは，一層困難であると考えられる．そこで，以下のように，システムが自動的にユーザの入力した質問を構造化することを考える．

情報検索においては，適合フィードバック(relevance feedback)[3]と呼ばれる方式が研究されてきた．適合フィードバックとは，検索結果に対してユーザが適合・不適合を判定し，その結果をもとに，質問の修正を行う方式である．通常の適合フィードバックにおいて，文書と質問はキーワードベクトルで表されており，キーワードの重み付けを変更することで質問の修正が行われる．ここでは，図 1 の様にページと質問がキーワードの階層構造で表されているモデルにおいて，質問の修正に適合フィードバックを用いることを考える．

具体的には，ユーザの入力した質問に含まれるキーワードから予め候補となる複数の構造化された質問例を生成し，それらが適合ドキュメント，非適合ドキュメントにどれだけマッチしたかを数えることで，最適な質問を選択する．

ユーザが入力したものと同じキーワードからなる質問でも，様々な構造を取ることができる．そこで，ユーザの入力したキーワードから，以下の手順で構造化された質問例を生成する．

1. 木における最大の節点数 N を指定する．
2. 節点数 N 以下のすべての木を生成する．
3. 各節点に一つ以上のキーワードを分配する．

最初に最大の節点数 N を指定するのは, 木の構造を組み合わせの数を制限することで，考慮しなければならない質問の候補を減らすためである．原理的にはいくらでも複雑な質問が考えられるが，Web に対してあまりに複雑な質問(例えば，4 階層以上の階層を持った質問)を用いても，合致するページは少なく実用性はあまりないと考えられる．

これらの質問例を用いて，以下のような方式で適合フィードバックを行うシステムを考える．

1. ユーザが複数のキーワード $k_1,k_2,\ldots,k_m$ で質問を投入する．
2. すべてのキーワードが含まれる Web ページを検索結果として返す. 同時にユーザの入力から複数の構造化された質問例 $q_1,q_2,\ldots,q_n$ を生成する．
3. ユーザが，適合ページ，非適合ページの判定を行う．
4. 構造化された各々の質問例 q_i の適合率 P_i，再現率 R_i を計算する．適合率は，質問にマッチしたページの中の適合ページの割合, 再現率は適合ページの中の質問にマッチしたページの割合である．
5. 適合率と再現率の調和平均[4]: $F_i = \frac{2}{\frac{1}{R_i}+\frac{1}{P_i}}$ などを評価値に用いて，質問例を一つ選択する．
6. 選択された質問例を用いて，検索を行う．

ここで調和平均は，適合率と再現率が両方とも高い質問を選択するために用いている．

5. 検索結果の比較呈示での応用

前節では，ユーザの要求に適した質問の構造化を推定する方法として，ユーザからのフィードバックにより得られた，適合ページ，非適合ページを用いる方法について提案した．これらは，ユーザがシステムと持続的なインタラクションを行うことを前提としており，ユーザにある程度の負担を強いる方法である．

しかし，予備実験の結果にみられるように，異なった構造の質問で大きく検索結果が異なる場合には，これらの質問の検索結果を対比しユーザに呈示することだけでも，ユーザが本当に欲しているものを発見することを支援できるのではないかと考えられる．

そこで以下のように，システムが複数の構造化された質問の中から，検索結果が大きくことなるものを選択してユーザに呈示する方法を考える．

1. ユーザが複数のキーワード $k_1,k_2,\ldots,k_m$ からなる質問を投入する．
2. 入力キーワードから, 複数の異なった構造化をされた質問 $q_1,q_2,\ldots,q_n$ を生成する．
3. 構造化された質問を検索エンジンに投入し, 質問ごとに検索結果の N 件ずつの URL リスト $U=\{L_1,L_2,\ldots, L_n\}$ を取得する．
4. URL の重複 $\left|L_{i_1} \cap L_{i_2} \cap \Lambda \cap L_{i_l}\right|$ を最小にするような l 個の URL リストを選択する．ここで l は予め与えられたパラメータである．
5. 選択された検索結果を並べてユーザに呈示する．

提案システムの処理の概略を図 2 に示す. 本システムでは,

通常の検索では混在して表示されてしまう検索結果を，構造化された質問ごとに分離して示すことで，ユーザが必要な情報を見つけやすくする効果を目指している．これは検索結果のクラスタリングと同様の効果があるのではないかと考えている．

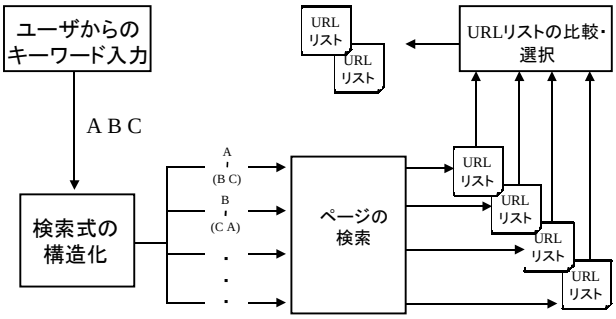

図 2　質問の構造化を利用した検索結果の比較呈示

Fig.2　Comparative Presentation of Search Results Using Query Structuring

図 3 に質問構造化システムのプロトタイプを示す．ここでは, Google が提供している Web サービスである Google Web APIs[2]を利用し，ユーザの入力した複数のキーワードを主題と本文の 2 階層に振り分け，検索結果の重複の最も少ない質問の組を呈示している．

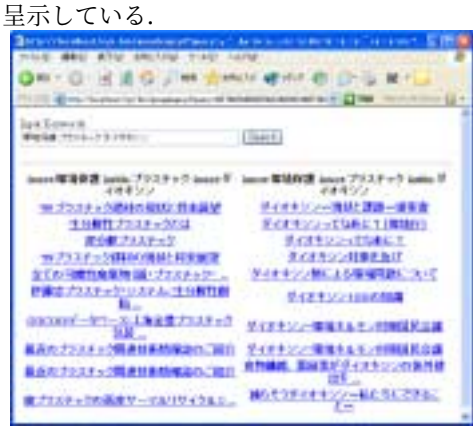

図 3　質問構造化システムのプロトタイプ

Fig.3　A Prototype of Query Structuring System

6. おわりに

本論文では, 既存の検索エンジンを利用した予備実験を行い，キーワードがページ内で現れる位置を考慮することにより，同じキーワードからなる質問であっても検索結果の内容が異なる場合があることを確認した．

その結果を踏まえて，Web における質問修正の新しい手法として，ユーザの入力した質問のキーワードを階層的に構造化して検索を行う手法を提案した．そのために，Web ページおよび質問をキーワードで木構造に表現し，両者の包含関係を利用して検索を行う方式を提案した．

また，ユーザからのフィードバックを利用し，質問に構造を与える方式についての検討を行った．さらに，複数の構造化された質問の検索結果の中で，なるべく異なるものを選択してユーザに同時に呈示することで，ユーザが意図したものを発見することを支援するシステムの提案を行い，プロトタイプを実装した．

今後は，見出しや表のタグなどを利用する方法や，[5]のように，単語の出現分布などを利用する方法を参考にし，Web ページから, キーワードの階層構造を抽出する方式について検討を行う予定である．また，大量の Web ページを対象にするためには，検索の高速化を行う必要がある．そのために，検索式とページのマッチングを高速に行うアルゴリズムの開発を行う．そして，実際の Web ページを収集して検索エンジンを構築し，提案手法の有効性の実証を行いたいと考えている．

[謝辞]

本研究の一部は，平成 14 年度科学研究費補助金特定領域研究(2) 「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」(課題番号 14019048, 研究代表者 田中克己) および基盤研究(A)(2) 「モバイル環境におけるコンテンツのマルチモーダル検索・呈示と放送コンテンツ生成」(課題番号 14208036, 研究代表者 田中克己) による．ここに記して謝意を表します．

[文献]

[1] Oyama, S., Kokubo, T., Ishida, T., Yamada, T. and Kitamura, Y.: Keyword Spices: A New Method for Building Domain-Specific Web Search Engines, Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01), pp.1457-1463 (2001).

[2] Butler, D.: Never trust a human, Nature, Vol.405, p.115 (2000).

[3] Salton, G. and Buckley, C.: Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science, Vol.41, No.4, pp.288-297 (1990).

[4] Shaw Jr., W.M., Burgin, R. and Howell, P.: Performance Standards and Evaluations in IR Test Collections: Cluster-Based Retrieval Models, Information Processing & Management, Vol.33, No.1, pp.1-14 (1997).

[5] 佐野綾一, 松倉健志, 波多野賢治, 田中克己: 部分グラフを基本単位とした Web 文書検索 : 単語の出現密度分布の適用, 情報処理学会研究報告, Vol.99, No.6199-DBS-119-44, pp.79-84 (1999).

小山 聡　Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助手．2002 京都大学大学院情報学研究科博士後期課程修了．博士 (情報学)．情報検索，データマイニングなどの研究に従事．電子情報通信学会，人工知能学会，ACM，AAAI 各会員．

田中 克己　Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 京都大学大学院工学研究科博士前期課程修了．工学博士．データベース，マルチメディアコンテンツ処理の研究に従事．情報処理学会，人工知能学会，日本ソフトウェア科学会，IEEE Computer Society，ACM 等各会員．

[2] http://www.google.com/apis/