

# 文書検索による概況調査支援システム

## Support System for Outline Investigation

### あらまし

文書データベースを対象としてユーザの概況調査(一つの事柄について様々な情報を収集し、整理して報告する)を支援することを目的としたシステムを紹介する。

本システムは文書データベースを対象に概況調査のための基礎データを収集し、必要な情報を抽出してユーザのニーズに沿う形に整形して出力する。ここではあらかじめ意味解析の結果を文書中にXMLタグの形で埋め込んでおき、検索コマンドで直接検索を行うことで、意味的な内容による高速な検索を実現し、さらに検索の失敗したときには、段階的に検索条件を緩和する機構を起動することで、意味解析の誤りや文書とユーザの質問における表現の違いなどに対応している。

### Abstract

This paper describes a system we have developed that supports outline investigations in text databases by providing essential information for making comparisons within retrieved data and for trend analyses of the data. In order to achieve high-precision semantic retrieval, which is often necessary for outline research, we added a semantic representation to the original text in the form of XML tags, which can be directly retrieved by our XML retrieval system. Each element of information is then extracted from the retrieved results and arranged in a well-organized table so that users can easily make further analyses.



伊吹 潤(いぶき じゅん)  
コンピュータシステム研究所ド  
キュメント処理研究部 所属  
現在、自然言語処理の研究開発に  
従事。



西野文人(にし の ふみひと)  
コンピュータシステム研究所ド  
キュメント処理研究部 所属  
現在、自然言語処理の研究開発に  
従事。



松井くにお(まつい くにお)  
コンピュータシステム研究所ド  
キュメント処理研究部 所属  
現在、自然言語処理、文書情報処  
理、情報検索の研究開発に従事。

## まえがき

我々が検索によって実際に問題を解決しようとする場合、検索結果として1件のデータを得るだけで済む場合はほとんどない。

例えば、購入対象として最適なデジタルカメラなどの製品の一つを選択したい場合、候補となる製品情報をいろいろと集めて比較することが必要だろう。また企業の信用調査を行う場合、その企業の関係した個々の出来事を集めた上で、その企業が信頼できるかを判断することになる。

著者らは上に述べたような概況調査（一つの事柄について様々な情報を収集し、整理し報告する）をサポートするためのシステムの開発を行っている<sup>1)</sup>

本稿では、概況調査におけるユーザ自身の作業を分析し、支援システムの設計目標と実装の方法について述べる。

## 概況調査における作業の分析

ここでは検索対象として、新聞記事のデータベースや企業Web上の公開ページを想定し、ユーザが製品調査や企業の動向調査を行う場合どのような作業が必要かを分析する。

### (1) 候補となる文書の収集

新聞記事やプレスリリースは出来事中心の情報源であるため、ユーザが特定のテーマのもとにまとめや概況調査を行うにはまず複数の文書を集めることが必要になる。この段階でどの程度正確な検索ができるかが後の作業に大きく影響する。製品調査や企業動向調査では、「出来事の主体」、「製品の機能」というように文書内の特定の部分についての条件を指定して文書を探したいことが多い。こうした条件をキーワードの組合せで表現するのは難しく、検索対象の文書がもともと多数なため、大量の検索ゴミが検索結果に入り込む危険性がある。

### (2) 情報の抽出と整理

検索ゴミの除去が終了した後も作業は残っている。文書群からユーザが必要とする情報を抽出して、それを整理された形にまとめて出力する作業が概況調査には不可欠である。

## 概況調査のための支援システムの設計

以上の分析から著者らの支援システムでは、ユーザ自身の概況調査に必要な情報を文献ベースから正確に検索

することと、ユーザの視点に沿った整理を行って出力できることを目標とした。以下、この二つをどう実現するかについて検討する。

対象文書を正確に検索するために

### (1) 意味的な内容によるタグ付け

文書の特定の一部に対しての条件指定を可能にするために、あらかじめ文書中に出来事の種類の（例：開発、販売）、構成要素（例：製品、主体企業）などに関する情報をXML形式のタグの形で埋め込んでおくこととした<sup>2)</sup> さらにXMLのタグごとに検索条件を指定できる検索エンジンを採用することによって意味的な内容に対する条件を利用して直接検索を行えるようにした。

### (2) 自然言語文による検索条件の指定

ユーザの検索要求は検索対象の文書と同じく、質問文の形で入力することとした。これは検索コマンドの形式を知らないユーザにも使えること、概況に対する検索要求を自然な形で表現できること、提示すべき情報の指定を質問文のトピックという形で同時に指定できるなどの利点のためである。

このため、本システムでは質問文を解析して、検索条件、質問文のトピックなどの情報を得るための枠組を用意した。

### (3) 検索結果を確保するために

意味検索による検索は一般にキーワード検索より条件が厳しくなるために検索ゴミの減少に大きな効果が見込まれる反面、検索結果が思うように集まらない場合が出てくる。

こうした場合、検索条件を修正して再検索をかけようとしても参考とすべき文書の情報が何もないため、再修正もままならない。

こうした状況を防ぐため、検索部に検索条件の段階的な緩和処理を導入して、1件以上の検索結果が得られるまで徐々に検索条件を緩めていくことにした。

検索の失敗の一因は、検索対象テキストのタグ付けの揺らぎや誤りだが、これらは特定のタグ（例：製品記述の部分）に対して起こることが多いので、タグ条件の緩和、あるいは単なるキーワード条件への置換えなどの操作によって条件を緩める。一方、ユーザの要求する種類の出来事がない場合は関連する出来事に展開する（例：販売記事がなければ開発記事へ展開）。

さらに緩和操作によっても検索結果が得られない場合は、一つひとつのタグに対する検索条件を単独で指定して検索を行ってみて、検索結果が0件となったものを失

敗原因としてユーザに伝える（例えば、企業に関する検索に失敗した場合、企業名、所在地などについての条件があればおのおのの条件単独で検索してみる）。

ユーザの視点に沿った情報の整理のために

検索結果から最終的な表示データを得る過程について説明する。

## (1) ユーザの求める情報の選択

まず質問のトピックをもとに抽出すべき項目を選択し、情報を抽出する。トピックだけでなく、関連するほかの項目（例：トピックが「製品」だったら「製造元」）も抽出対象とする。

## (2) 情報の統合と提示形式の決定

抽出した情報について、どのように見せるべきかを決定する。1件の情報についてどう見せるか（各項目をどのコラムに配置するか）については基本的には情報をユーザの関心の度合に従って並べて配置する。

各情報をどのように配置するかについてはトピック項目をキーとしてソートする。また必要に応じてサマリー的な情報を付加して、情報の概観が得られるようにする。

## システム全体の構成

システム全体での処理の流れを図-1に示す。

ユーザの質問はまず意味解析によって意味表現と質問のトピックに変換される。意味表現は検索条件に変換されて文書データベースに対する検索が行われる。検索が失敗した場合には、条件緩和操作が起動され、検索が成功するまで検索条件は徐々に制約の緩いもの書き換えられる。

検索に成功すれば、検索結果からは質問のトピックによって指定される項目の情報が抽出され、さらに加工されて表形式の概況情報として出力される。

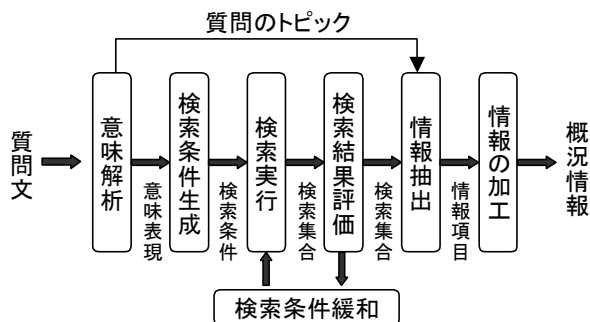


図-1 処理の流れ

Fig.1-Flow chart of processing.

## 質問の処理例

ここでは「Macの上で動く占いソフトを知りたい」という質問文例に対する処理を順に説明する。

### (1) 質問文の解析

意味解析の結果は次のようになる。「製品条件部」に示されているのは製品に対する検索条件であり、キーワードとしてMacや占いがマークされている。また「q="製品"」の部分は質問文のトピックが製品にあることを示す。

<検索要求 q="製品">

<製品条件部>

<KW>Mac</KW>の上で動く<KW>占い</KW>

</>

<製品種>ソフト</>

<要求部>を知りたい</>

</>

### (2) 検索条件の生成と実行

ここでは検索に失敗して、検索条件の緩和が行われた例を示す。

- ・検索条件生成部で、つぎの条件が生成される。

条件1: SELECT 販売情報 WHERE <製品種>Mac&占い&ソフト</>

- ・検索結果が0件だったために、検索失敗と判断。

- ・検索条件緩和処理を起動して次の条件を生成

条件2: SELECT 販売情報 WHERE <kw>Mac</kw>&<製品種>占い&ソフト</>

(「製品種」タグに対する「Mac」の指定からタグ名なしで「Mac」を指定するキーワード検索に条件を緩和)

- ・2件が検索され、検索成功で終了する。

検索失敗の原因の一つは意味解析の失敗である。例えば、ソフトの動作環境に関する情報は機能などの基本的な情報と違い、別の文に書かれることが多い。

このため、解析には文脈の処理が必要であり、失敗して何のタグも付かない状態になることが多い。

### (3) 検索情報の整理と提示

質問文に対する最終的な表示結果を図-2に示す。ここでの表示をどう決めたかについて具体的に説明する。

- ・まず、質問文のトピックが「製品」なので、この例では「製品の種類」、さらに意味モデル上で関連性リンクを持つ「企業」(表示結果では販売元)の二つを提示対象として選ぶ。

## 実行結果のサマリ

2件の記事が検索されました

## 実行結果詳細

製品の種類	名称	販売元	
任天堂の携帯型ゲーム機	ゲームボーイ	イマジニア	本文
パソコン上でタロット占いができるソフト	バーチャル・タロット	エー・アイ・ソフト	本文

図-2 検索結果の表示形式  
Fig.2-Format of retrieval result.

- ・ つぎに対象物の属性からどれを表示するかを決める。  
例えば、製品はトピック項目なので詳しく表示し（製品の種類、名称など）、企業の方は関連項目なので一つ（名前）だけを表示する。
- ・ おおのこの項目に質問文中の言葉を利用して見出しをつける（例えば、質問文中に「企業」という言葉があれば「組織体名」を「企業名」にする）。
- ・ 最後に表中の各行また、同一種類の製品が近くに来るように情報を製品種によってソートする。

## システムの評価結果

日刊工業新聞8年分の記事をデータベースに登録し、企業の経済活動に関する質問20文を収集して実際に検索を行ってみた結果について述べる。

検索の適合率は平均して約8割であり、キーワード検索の最高で3割の適合率から大きく改善されている。その際の再現率はキーワード検索結果と比較して7, 8割であった。

また表形式の結果表示は、データ件数が多い場合や、長い文書の場合の分析作業に有効であり、対象にあいまいさがある場合（例：キーボード）にも表示データのソートキーをうまく選択することによって必要な部分のみの抽出をより簡単に行うことができることを確認した。

## む す び

以上のことから限られた対象ではあるが、本システムが概況調査のために不可欠な作業（必要なデータの選別、データ相互の比較）を能率化するのに有効であり、今後の開発のプロトタイプとして使える見通しを得た。

インターネットの普及により出来事や科学データなどの一次情報の蓄積が進んでいるが、概況的な情報は株式市況などの限られた分野で人手で作成されている状況である<sup>(3)</sup>。

著者らは今後、とくに様々な情報源や分野に有効に対処するためにシステムの持つ解析・検索機能に埋め込まれた知識の構造化、知識を獲得するためのツール類の整備を重点的な目標としている。

## 参 考 文 献

- (1) 伊吹潤ほか：質問文からの検索条件と提示情報の決定．言語処理学会第6回年次大会，2001．
- (2) 西野文人ほか：新聞記事からの人物・企業情報の抽出．情報処理学会研究報告，NL127-17，p.125-132（1998）．
- (3) （財）データベース振興センター編：データベース白書 2000．