

ユーザの思考を妨げないペン操作によるインタラクティブな情報検索

～ 意味解析と意図推定に基づく連鎖情報検索 ～

鈴木優 布目光生 石谷康人

(株) 東芝 研究開発センター 知識メディアラボラトリー

1 はじめに

情報検索は、文書作成や計画立案などの知的生産を行う過程で頻繁に必要とされる行為である。それ故、情報検索のプロセスが煩雑であると、ユーザの思考が妨げられ、知的生産が阻害されてしまう恐れがある。現状の情報検索では、ユーザは、(1) 検索クエリの検討、(2) 検索クエリの入力、(3) システムが提示した検索結果の吟味、(4) 目的に合致する情報の選択および閲覧といった作業をキーボードとマウスを併用しながら実施している。これらの作業は必ずしも簡便であるとは言えず、ユーザは情報検索の度に知的生産を中断しなければならないといった問題点があった。そこで、本論文では、ユーザの思考や知的生産を妨げることなく、閲覧中の文書に関連する情報を簡単な操作で的確に見つけることができる新しい情報検索方法を提案する。

2 基本コンセプト

紙媒体の文書を読んでいる時、意味が分からない語句や重要な語句に出会うと、その箇所にアンダーラインを引くことが多い。このような行為は多くの人々が長年にわたって慣れ親しんだものであり、その度に思考が中断されるようなことはない。そこで本論文では、閲覧中の文書の任意の語句にアンダーラインを引くと、その語句に関連する情報を検索するユーザインタフェースを備えた情報検索方法を提案する。

情報検索では大抵の場合、検索結果が順序付けられて出力される。このとき、所望の情報が必ず上位に出力されるようになっていれば、情報アクセスのスピードを格段に速くすることができる。このように、情報検索において検索結果を的確に絞り込むことができれば、思考や作業の中断を回避することができる。そこで本論文では、ユーザとのインタラクションによってユーザの検索意図を推定しながら検索結果を的確に絞り込むことを可能とする検索方法についても併せて提案する。

以下に、提案方法における情報検索の手順について述べる。

Step 1: 閲覧中の文書において、ユーザがキーワードにアンダーラインを引く

Step 2: キーワードとその周辺の語句の意味から、システムがユーザの検索意図を推定し、検索方法の候補を提示する

Step 3: 提示された検索方法の候補から、ユーザが適切な検索方法を選択する

Step 4: ユーザが選択した検索方法に基づいてシステムが検索処理を実行する

Step 5: システムが出力した検索結果から、ユーザが所望の情報を閲覧する

Interactive and Successive Information Retrieval with a Pen-based Interface
Masaru SUZUKI, Kosei FUME, Yasuto ISHITANI.
Toshiba Corporation.
masaru1.suzuki@toshiba.co.jp

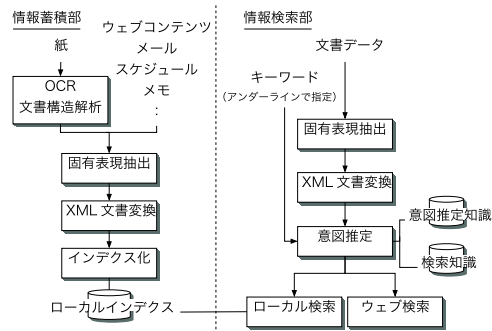


図 1: 情報検索システムの構成図

情報検索では一般的に、検索結果を的確に絞り込むためには検索キーワードを上手に設定する必要がある。提案方式では、このような検索キーワードの設定を上述した手順に基づくマンマシンインタラクションによって可能としている。この結果、「検索キーワードの簡単な設定」と「的確な絞り込み検索」の両立が可能となるため、ユーザの思考や作業を中断しない情報検索を実現することが可能となる。

3 情報検索システムの構成

本論文で提案する情報検索システムは「情報蓄積部」と「情報検索部」で構成されている(図 1)。以下に、情報蓄積部と情報検索部の構成について概説する。

3.1 情報蓄積部

OCR・文書構造解析: 紙文書が入力された場合、文献 [1] に基づく OCR 処理と文書構造解析処理を適用することにより、紙文書をプレーンテキストもしくは章節構造や箇条書き構造を保持した XHTML 文書に変換する。

固有表現抽出: 入力文書に対して文献 [2] の固有表現抽出処理を適用することにより、文書中の語句に対して人物 (PERSON)、会社名 (COMPANY)、E-mail (EMAIL)、URL (URL)、場所 (JAPAN_CITY)、日時 (DATE, TIME)、数量 (NUM)、金額 (MONEY) などの固有表現クラスを XML タグとして付与する。固有表現抽出結果の例を図 2 に示す。

XML 文書変換: 固有表現抽出結果に対して文献 [3] の XML 文書変換処理を適用することにより、固有表現のグループやグループの階層構造などを抽出して XML 化する。

インデックス化: 入力文書に対して、形態素単位のインデックス [4]、固有表現とクラス情報のインデックス、固有表現のグループや階層構造のインデックス、文書の種別などの書誌情報のインデックスをそれぞれ作成する。

3.2 情報検索部

固有表現抽出, XML 文書変換 文書データを閲覧する際に、バックグラウンドで文書から固有表現を抽出すると共に固

<ORG>技術委員会</ORG>の開催について\\<EVENT>標記会議</EVENT>を下記により開催いたしますので、万障お繰り合わせのうえご出席下さいませよう、ご案内申し上げます。\\記\\日時:<DATE>平成17年3月7日</DATE>(月)<TIME>午前10時</TIME>~<TIME>11時30分</TIME>まで\\場所:(株)<COMPANY>東芝</COMPANY>(<JAPAN_CITY>東京都港区芝浦</JAPAN_CITY>1-1-1) 本社ビル<NUM>39階</NUM>大会議室\\議題:インタラクシオン2005参加報告/<PERSON>鈴木優</PERSON>(<ORG><COMPANY>東芝</COMPANY>研究開発センター</ORG>) \\以上

図 2: 固有表現抽出結果の例

有表現のグループおよびグループの階層構造を抽出する。さらに処理結果をXMLで記述する。

意図推定 意図推定部では、まず、文書中のアンダーラインが引かれている語句(キーワード)と、キーワードの固有表現クラスと、キーワードが含まれるグループ内に存在する他の固有表現クラスを順次抽出する。次に、これらの情報を用いて図1の意図推定知識を参照することにより検索方法の候補を出力する。検索方法の候補には、「地図検索」「メール検索」「人物検索」「住所録検索」「商品検索」などがあり、一度に複数の候補が出力される場合にはユーザによって適切な検索方法が選択される。各検索方法では、ウェブ検索/ローカル検索の指定、検索クエリの拡張を目的としたXPath形式の階層情報、検索方法を表す定型語句(例:“地図”), 検索対象の種別(インターネット上のWebサイトの種別を含む)などが検索知識として定義されており、これらは後段で絞り込み検索を実現するためにウェブ検索やローカル検索で活用される。

ウェブ検索 ユーザが選択した検索方法の検索知識において「ウェブ検索」が定義されている場合には、キーワードに階層情報や定型語句など追加した拡張クエリを用いて情報検索サイトを介した情報検索を行って絞り込み検索を実施する。

ローカル検索 ユーザが選択した検索方法の検索知識において「ローカル検索」が定義されている場合には、ウェブ検索と同様に検索クエリを拡張すると共に、検索対象の種別に関する検索知識を用いて絞り込み検索を行う。このとき、ローカルインデックスを参照することにより、メール、スケジュール、メモ、紙文書などの文書種別の絞り込み検索、人物、会社名、場所、日時などの固有表現クラスによる絞り込み検索、グループの階層構造による絞り込み検索などを実施できる。

4 インタラクティブな情報検索の例

以下では、出張計画を立案する例を用いてインタラクティブな情報検索のユースケースについて説明する。ここでは、出張計画を立案する際にスケジュール情報に事前登録されているイベントに関連する情報をデスクトップやインターネットから検索する場合を想定する。まず、スケジュール情報のイベントの箇所にアンダーラインが引かれると、その前後にある日付情報や時間情報を用いて「メール検索」や「スケジュール検索」などの検索方法が提示される(図3)。このとき、ユーザがメール検索を選択するとデスクトップにある当該イベントの開催案内メール(図4)が検索される。次に、開催案内メールに記載されている所在地の箇所にアンダーラインが引かれれば(図4(a)), そこに付与されている固有表現クラス(図2の<JAPAN_CITY>)を用いて「地図検索」という検索方法が提示される(図4(a))。また、開催案内メールに記載されている人物名の箇所にアンダーラインが引かれれば(図4(b)),

3/7(月)	10:00-11:30	技術委員会	
3/8(火)	[追加]		メール検索 スケジュール検索
3/9(水)	10:00-17:00	テーマミーテ	
3/10(木)	12:00-17:00	成果発表会	ウェブ検索 デスクトップ検索
3/11(金)	10:00-12:30	委員会	デスクトップ検索 辞書検索
3/12(土)	[追加]		
3/13(日)	[追加]		

図 3: スケジュール閲覧中の検索例

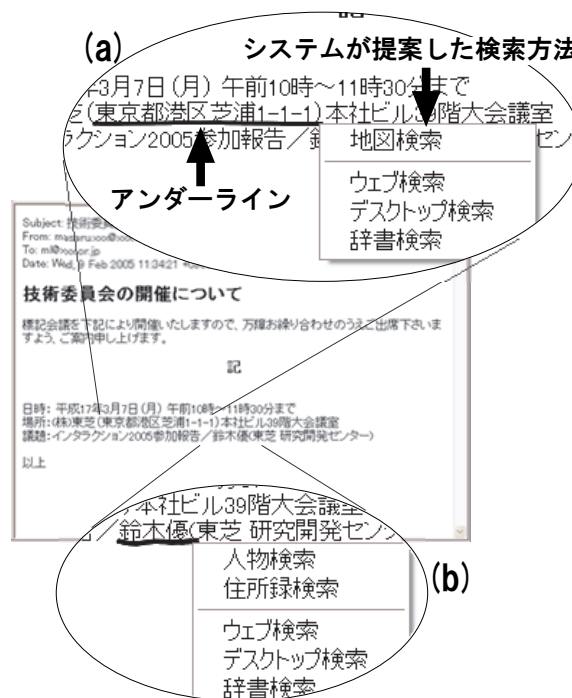


図 4: メール閲覧中の検索例

そこに付与されている固有表現クラス(図2の<PERSON>)を用いて「人物検索」や「住所録検索」という検索方法が提示される(図4(b)). 例えば地図検索では、検索クエリとして、所在地情報(図2の「東京都芝浦1-1-1」と、定型語句「地図」と、所在地の周辺にある会社名(図2の「東芝」)を用いてWeb検索が行われ、その結果、所望の地図情報を的確に検索することが可能となる。提案方法では、このようにユーザとシステムとのインタラクションにより、簡単なペン操作だけで関連する情報に連鎖的にアクセスすることを可能となる。

参考文献

- [1] 石谷康人 他. 紙文書を対象としたピボットXML文書に基づくXML文書変換システム. 電子情報通信学会 技術研究報告, PRMU2003-216:7-12, 2004.
- [2] 市村由美 他. 質問応答と、日本語固有表現抽出および固有表現体系の関係についての考察. 情報処理学会 研究報告, NL-161-3, 2004.
- [3] 布目光生 他. 表層表現抽出と文書構造解析に基づくXML文書変換システム. 情報処理学会 研究報告, 2004-DD-46:1-8, 2004.
- [4] (株) 東芝. <http://pf.toshiba-sol.co.jp/prod/km2/index.j.htm>.