検索結果を精度よく絞り込むための類似検索システムの開発
—“選ぶだけ”で“見える”絞り込みキーワードの提示システム—

1. 背景

インターネット技術の普及や計算機性能の向上に伴い、今までは紙を媒体として配布されていた文書が電子化されて公開、配布される機会が増大している。そうした大量のデータから目的とする情報を探し出す手段として、情報検索システムが普及しており、World Wide Web（以降、Web）における様々なWeb 検索サービスとしてGoogle 等のサービスは、特に身近な存在となっている。しかし、最も古典的な方法であるキーワード型検索方法においてユーザは求める文書を得るために検索式（クエリ）を作成する必要があるが、検索意図を正確に表現した検索式を作成することは難しい。その弱点を補う従来型の「類似検索」と「選択式のクエリ拡張技術」においても以下の問題点がある。

(1) 類似文書として指定する文書が少ない場合に絞り込みキーワードを確定できない。
(2) 絞り込みキーワードの中から適切な追加キーワードを選択するのが困難である。

このためにユーザの求める結果を得ることができず、ユーザにとっての負担が大きかった。

2. 目的

本プロジェクトでは、従来の二つの問題を解決するために複数の正解／非正解文書を基にした検索式の生成による絞り込み支援法を実現する。

ユーザに検索意図に近い文書または外れている文書を指定させることにより、ユーザの検索意図を類推する。この類推結果を基にキーワードを抽出し提示することで、検索式の作成が容易になる。

さらに、ユーザがキーワードを選択すると実際に絞り込む前に検索結果から除外される文書を視覚的に表示する機能により、検索意図に対して妥当な検索式なのか確認することが可能となる。

これらにより、既存の検索エンジンにおいて所望の文書を簡単に精度よく得ることを実現する。

3. 開発の内容

上記目的を実現するクライアント・サーバ構成のシステム一式を開発した。本システムのクライアントソフトウェアはWeb ブラウザのFirefox 上で任意のページでユーザの作成したJavaScript を動作可能とする追加拡張機能の「Greasemonkey」のユーザスクリプトとしてインストールする事で動作する。

ユーザが既存の検索エンジンサイトで検索を行う際に、検索キーワードを入力し検索を行った後の検索結果の一覧表示画面において、クライアントは自動的にサーバとの通信を行い、各々の文書の候補キーワードを取得する。

次の図１に示すように、各々の文書について、その文書を正解文書（または非正解文書）として絞り込む為の候補キーワードを表示する。また、キーワードの上にマウスカーソルを合わせる事で、絞り込まれる文書がハイライト表示される。

また、チェックボックスを選択するとそれらの文書を正解／非正解文書としてグルーピングしたキーワードが表示される。

図 1：クライアント動作画面

サーバ側はクライアントからの通信をトリガーとして動作し、ユーザの指定した複数の正解／非正解文書から検索式の抽出を行う。サーバ側は組み込み型の Web サービスとして動作するよう実装している。サーバ側で実装している処理は次に示すクローラ・キーワード抽出・スコアリング機能である。

(1) クローラ（本文抽出）
本システムでは、クローラはクライアント部から送信されてきた検索結果の一覧から各文書の本文を取得し、HTML タグの除去を行って正規化を実施している。この際、埋め込み JavaScript も併せて除去する機能を実装した。

(2) キーワード抽出
(1) で抽出した本文から検索式の候補となるキーワードを抽出し、正解文書に含まれるキーワードリスト、非正解文書に含まれるキーワードリストなどから次の検索式候補となるキーワードをリストアップする機能を実装した。

(3) スコアリング処理
(2) で抽出したキーワードのリストから各キーワードの正解／非正解文書中の出現頻度を計測し、重みづけを行い、有意なキーワードを選定する機能を実装した。この重みづけは正解文書にのみ出現するキーワードや、非正解文書のみに出現するキーワードでなおかつ、多くの文書には含まれないキーワードにより高い重みをつけるよう実装した。
この重みに基づいて、より重みの高いキーワードがユーザの絞り込みを支援する

ための次の検索式候補として、クライアント部へ送信するよう実装した。

4. 従来の技術（または機能）との相違

本システムはユーザからの明示的なフィードバックを受け取り、検索式（クエリ）の書き換えをおこなうシステムである。このようなシステムは適合フィードバックシステムと呼ばれ、これまで多くの研究がなされてきた。

しかし、フィードバックに用いる文書数が少ない場合には重要なキーワードを判断できないためどうしても精度が落ちてしまう問題があった。その対策として本プロジェクト同様にユーザのキーワードの選択によるクエリ拡張をおこなうシステムもあったが、キーワードと文書との対応がわからないので適切なキーワードを選択するのが困難という問題があった。また、多くのシステムが特殊な環境や専用の検索エンジンでのみ実行され、Google などの一般的な検索エンジンの結果を利用することは難しかった。

本システムでは、文書数が少ない場合の精度の問題をユーザのキーワード選択により解決し、さらにキーワードと文書との対応を効率的に確認することができる。さらに、様々な検索エンジンやサイトの検索機能に対応することは実用性の面で大きなアドバンテージであると考える。

5. 期待される効果

本システムの効果として以下の 5 つが考えられる

i. 正解文書が見つかれば回覧行為が無駄にならず、検索時間が収束しやすい。
ii. 類似文書群を数多く得られるので、先行例を探すような検索（裁判判例，学術論文，特許，レシピ）において利便性が高い。
iii. 既存のさまざまな検索サービスと組み合わせて類似検索可能。
iv. 文字入力が PC に比べて不自由なモバイル端末でも選択する操作で絞り込むので利便性が高い。また、検索スキルが低くとも適切な検索式を構築できる。
v. 検索者の意図を把握しやすく、的確な広告を出せる可能性がある。

6. 普及（または活用）の見通し

本システムは主に 2 通りの利用方法が考えられる。

i. プロジェクト期間内に実現した、ブラウザのプラグインとしての利用方法
ii. サイトの検索機能に埋め込む形での利用方法

ⅰ.に関しての利点は検索対象が公開されていればどのようなサイトや検索エンジンでも本システムの類似検索を利用できる可能性がある。その反面、ユーザが検索する対象範囲が広いので高速化に必要な事前の文書解析（インデックス生成）をおこなうのに労力がかかる、ユーザはプラグインのインストールが必要などのデメリットが存在する。

ⅱ.に関してはニュースサイトやブログ、ショッピングサイトなど検索機能を持っているサイトに、クライアントを埋め込む利用方法である。対象サイトの協力が欠かせないが、代わりに検索対象の範囲が限定できるので高速化が容易、ユーザがプラグインのインストールを必要としないなどの利点がある。

実際の展開にあたっては、ⅰ.とⅱ.の利用を両輪で進めていく。ⅰ.のサービス公開により、評判を上げ、それを実績にⅱ.の導入を進める。さらにⅱ.の導入を進めること

で、ｉ.で利用するインデックスの規模を拡大し性能を向上させてゆく。
とくに電化製品などのショッピングサイトや旅行サイトなどは類似の商品が結果として表示されることが多く、多数の属性により目的の情報を絞り込む必要があるが、「“選ぶだけ”で“見える”」機能に対応することで利便性の向上が大きく見込まれる。

7. クリエータ名（所属）

有澤　悠紀（キヤノン株式会社）
大西　雄一朗（株式会社コンピュータシステムエンジニアリング）