IPA

2010 年度未踏 IT 人材発掘・育成事業　採択案件評価書

1. 担当 PM

平本　健二 PM（経済産業省　CIO 補佐官）

2. 採択者氏名

チーフクリエータ：有澤　悠紀（キヤノン株式会社　ソフトウェア基盤第一開発部
ソフトウェア基盤 12 開発室）
コクリエータ：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
技術本部品質保証部品質保証第二課）

3. 委託金支払額

2,880,000 円

4. テーマ名

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

5. 関連 Web サイト

なし

6. テーマ概要

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。

そこで本プロジェクトでは検索結果を精度よく絞り込むための類似検索システムを開発した。本プロジェクトの特徴はユーザからのフィードバックによる検索キーワードの提示を行う点である。本プロジェクトの成果では、ユーザの指定した文書から絞り込むキーワードの抽出と提示を行い、さらにユーザがキーワードを選択すると検索結果から除外される文書を表示することが可能となった。これによりユーザは、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ができるようになる。

7. 採択理由

検索の精度を上げるために、類似検索から絞り込みを行うのではなく、正解文書と非正解文書をキーにして文書検索精度を上げていくアプローチは、実際の検索者の意識に近く、従来の検索のように絞り込み過程で正解文書まで検索対象外にしてしまう恐れも少ない。政府内のホームページの検索は、不要な文書が表示されると評判がわるく、そこを改善する手段として有効と考えられる。

また、選択により検索対象を絞っていく方法であり、将来はグローバルに対応できるところも評価ができた。

8. 開発目標

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。

9. 進捗概要

本プロジェクトの目標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。

10. プロジェクト評価

検索者の「意思」を重視して検索を高度化しようというアイデアは非常によい。また、プロジェクト実施にあたり、性能が出ない場合には次々と新しいツールを導入し検証してプロジェクトを進めていたことは評価できる。また、正解文書と非正解文書で選択したときに直接検索クエリにするのではなく、候補単語を選択することで検索結果のプレビューができる工夫などユーザインタフェース面で工夫を行ったことも評価できる。しかしながら、実装に時間が掛かり、プロジェクト全体が遅れたことは否定できない。短時間で検証し実装する技術力を磨くことが重要である。また、プロジェクト途中で製品選択などの応用利用の領域に行ってしまい方針がぶれたのも計画が遅延した原因と考えられる。新しいことにチャレンジすることと、着実に納期に完成させることを両立できるよう、プロジェクト管理に取り組む必要がある。

11. 今後の課題

今後の課題としては、プロジェクト期間中にできなかった、ランキング処理のチューニングによる高速化、ブラウザの追加拡張機能の配布がある。また、プロジェクト中にも実施した辞書の整備は定期的に実施する必要がある。