社内文書検索システム（1）
-検索プラットフォーム CRISP-

白石 展久†
NEC サービスプラットフォーム研究所†

1 はじめに

企業内の情報には、情報間のリンクが少ない、十分にキーワードを含んでいないなどの問題があり、社内文書検索システムでは、これらの問題を解決する、利用目的に応じた柔軟なスコアリングが求められる。また、情報が複数のシステムに分散して蓄積されているため、社内文書検索システムはそれらと柔軟に連携する必要がある。本稿では、利用目的に適した検索サービスを実現する検索プラットフォーム CRISP を紹介し、その以下の特長:

(1) 検索システムのエンジン構成を動的に変更可能な分散アーキテクチャ

(2) 検索目的に応じたエンジン群の自動的な重み付けを実現するソムリエエンジン

について述べる。

2 CRISP のシステム構成

検索プラットフォーム CRISP のシステム構成を図 1 に示す。CRISP は、検索システムにおけるエンジン群(スコアリングエンジンや検索エンジン)の構成を動的に変更可能とする分散アーキテクチャとなっており、利用目的に応じて適切なエンジン群を選択し重み付けを行うソムリエエンジンを持つ。

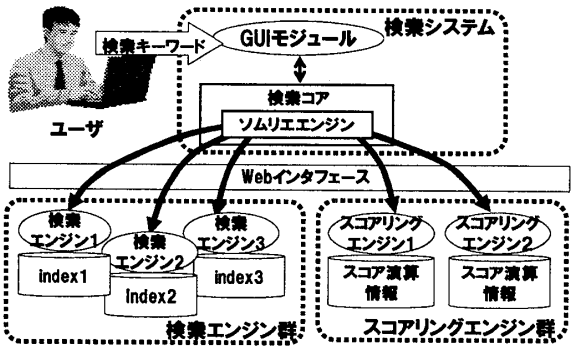

図 1: CRISP のアーキテクチャ

An In-house Document Retrieval System (1) –Search Platform "CRISP" -

† Nobuhisa Shiraishi, Service Platform Research Laboratories, NEC Corporation.

CRISP は、検索要求を受け取ると、接続されている各検索エンジンに並列に検索要求を行う。そして各検索エンジンから検索結果応答を受け取ると、その検索結果リストを集約し、検索システムに接続されている各スコアリングエンジンへ、並列にスコア演算要求を行う。そして各スコアリングエンジンからスコア演算結果を受け取ると、後述のソムリエエンジンによるスコアリングエンジンの重み値情報を元に検索結果のランキング調整を行い、ユーザに対して最終的な検索結果リストを表示する。

3 分散アーキテクチャ

3.1 特長

CRISP は、検索システムにおけるエンジン群(スコアリングエンジンや検索エンジン)の構成を動的に変更可能な分散アーキテクチャとなっている。これにより以下の効果がある。

(1) 横断検索の対象システム範囲の拡張促進

CRISP に接続する検索エンジンは、後述のエンジン接続インタフェースに従って自由に開発することができる。例えば、企業内の既存情報システムの運用部門は、そのシステムに格納された情報を検索する検索エンジンを、独自に開発することができる。更に新規に開発された検索エンジンは、そのアクセス URL を後述のソムリエエンジンに登録することにより、検索システムのサービスを停止させることなく動的に組み入れることが可能であり、検索システムの複数の既存システムを跨いだ横断検索の対象システム範囲の拡張が促進される。

(2) 検索システムの開発者と利用者によるスコアアルゴリズムの拡張促進

検索エンジンと同様に、スコアリングエンジンも前述のエンジン接続インタフェースに従って自由に開発することができる。冒頭で述べた通り、企業内の情報はインターネットの情報とは性質が異なるため、tf*idf アルゴリズムや PageRank アルゴリズムのようなインターネット検索において有効なスコアアルゴリズムとは異なるアルゴリズムを適用する必要があり、CRISP の分散アーキテクチャは、このような様々なスコアアルゴリズムを検索システムに組

み込むのに非常に適している。CRISP では、検索システムの開発者のみでなく、その利用者も様々なスコアアルゴリズムをスコアリングエンジンとして実装し、自分が使用する検索システムに自分が実装したスコアアルゴリズムを動的に組み入れることができる。

3.2 エンジン接続インタフェース

CRISP では、検索システムを構成するエンジン群は Web インタフェースで接続される。各検索エンジンやスコアリングエンジンは、検索要求を格納した RDF/XML を POST リクエストで受け取り、検索結果を格納した RDF/XML をレスポンスとして戻す。RDF/XML によって検索要求と検索結果をやり取りすることにより、検索システムから検索エンジンへの検索要求に任意のメタ情報を容易に追加可能であり、また各検索エンジンからの検索結果の文書情報にも任意のメタ情報を容易に追加可能である。このため、検索システムからの検索要求にメタ情報を追加することによる検索エンジンの検索精度の強化や、検索エンジンからの検索結果の文書情報に追加されたメタ情報を処理表示する機能を追加することによる検索結果の表示内容の充実が容易に可能であり、検索システムの機能拡張性が高い。

4 ソムリエエンジン

4.1 特長

CRISP では検索システムの開発者のみでなく、利用者も検索エンジンやスコアリングエンジンを実装して検索システムに動的に組み入れることが可能であるが、検索システムの拡張が促進される反面、様々なエンジン群が自由に動的に検索システムに組み込まれるため、検索システムには、ユーザやユースケースに応じて自動的にエンジン群の重み付けを行う機構が必要となる。CRISP の「ソムリエエンジン」は、ユーザの検索結果クリックを元にエンジン重み値を算出し、検索実行時に各エンジンの重み値を検索結果文書のスコアに反映させることにより、検索システムは検索時に各ユーザに最適なエンジン構成を動的に構成し、各ユーザの嗜好にあったスコア算出を行うことができる。また、検索キーワードごとにエンジンの重み値を導出して蓄積することにより、検索キーワードやそのカテゴリに応じた最適なエンジン構成を動的に構築することも可能になる。

4.2 エンジン重み値の算出アルゴリズム

ソムリエエンジンは、ユーザの検索結果文書のクリックイベントを収集管理し、クリックされた文書のスコア情報から、スコアリングエンジンの重み値を導出する。スコアリングエンジンの重み値は、ユーザがクリックした文書に対して当該スコアリングエンジンが算出したスコアの、総合スコアに対する割合の累積を用いる。ユーザが検索結果のある文書をクリックした場合、その文書に対するスコアリングエンジン A の算出スコアを Sa、その文書の総合スコアを S とすると、スコアリングエンジン A の重み値 Wa は、スコアリングエンジン A の文書のクリック前の重み値を Wi、Wi の最終更新時刻から現在までの経過時間を t、経過時間による重み値の減衰率を r として、以下の式で算出される。

$$Wa = Wi \times t \times r + \frac{Sa}{S}$$

重み値 Wi を時間減衰させることにより、直近のスコアリングエンジンの重み値が、重み値により強く反映される。なお、スコアリングエンジンの重み値の初期値は 1 とする。

4.3 エンジン重み値の文書スコアへの反映

検索実行時、CRISP は検索要求を受け取ると、前述の方式によって算出・蓄積した各エンジンの重み値によって、各スコアリングエンジンが算出したスコアを重み付けして足し合わせ、文書の総合スコアを算出する。なお、文書スコアの算出における各スコアリングエンジンの算出スコア値とエンジン重み値との演算は、単純な行列積算アルゴリズムのみでなく、様々な演算アルゴリズムの実装を準備することによって、ユーザ情報やユースケース情報等を元に、ソムリエエンジンが演算アルゴリズムを指定することも可能である。

5 おわりに

本稿では、検索プラットフォーム CRISP を紹介し、Web インタフェースによる分散アーキテクチャとその利点、ソムリエエンジンのユーザの検索結果クリック行動情報を元にしたスコアリングエンジンの重み付け方式とその利点について述べた。現在、筆者らは CRISP アーキテクチャによる社内文書検索システムを社内で実験公開しており、その社内文書検索システム上でソムリエエンジンを試作して動作させようとしている。今後はこの CRISP ベースの社内文書検索システムに接続するスコアリングエンジンや検索エンジンを充実させながら、ソムリエエンジンの効果を測定し、考察と改良を進めていく。