分散協調サーチエンジンにおけるスコアリング

3T－4

佐藤永欣 山本崇 西田喜裕 上原稔 森秀樹†
東洋大学工学部情報工学科‡

1 はじめに

集中型サーチエンジンでは、大量の文書の収集やインデックスの管理が問題になる [7][8]。そこで、分散したサーチエンジンが協調して検索を行う協調サーチエンジン (Cooperative Search Engine,CSE) を提案した。あるサーチエンジンに情報が無くても、情報を持っている他のサーチエンジンを教えてもらうことで間接的に検索できる。CSE では分散した多数のサーチエンジンや、何種類かのサーチエンジンを利用可能だが、正しい検索を行うためには検索時のスコアの扱いが重要になる。本稿では CSE でのスコアの扱いについて述べる。

2 CSE の構成

CSE には二つの構成要素がある。一つは文書の検索、ユーザーとの対話を行う Local Search Engine(LSE)、もう一つは与えられたキーワードについて詳しい情報を持っている LSE を捜す Location Server(LS) である。CSE ではこれらの構成要素が協調して動作することにより分散した文書の検索を可能にしている。LSE はいくつ存在してもよく、通常は Web サーバ一つについて LSE が一つ存在する。しかし、複数の LSE が一つの Web サーバに同居したり、他の Web サーバの情報を持つことも可能である。これらの LSE は LSE 同士、または LS と通信しながら検索を行う。CSE のこのような動作モデルは以下のようなメリットを持つ。

- 集中型サーチエンジンではインデックスの更新を行う毎に各 Web サーバが持つ文書を転送する必要があるが、CSE では LS にキーワードとスコアの情報を転送すればよいだけなので更新時のトラフィックの軽減が期待できる。
- 各 Web サーバに検索機能を持たせることによりサーチエンジン専用のマシンが不要になる。つまり検索負荷を分散できる。
- 組織のある部門に関する情報はその部門が管理する、という情報の扱い方が理想的だが、各部門の Web サーバが自分の持つ情報について検索すれば、この理想に一歩近付くことができる。

3 CSE の動作

CSE の動作には大きく分けて検索時と更新時の動作がある。Fig.1 に CSE の検索時の動作の概略図を示す。

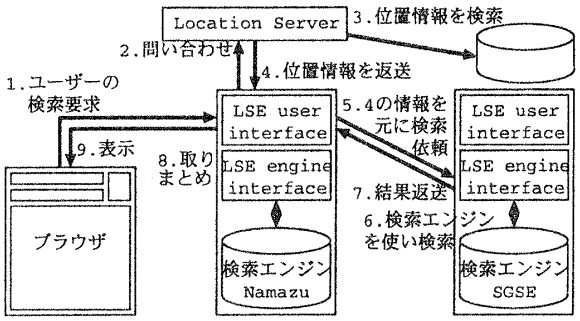

Fig. 1. CSE の全体的な動作の概念図

ユーザーの検索要求は入力フォーム付のページに対する入力としてブラウザ等から LSE に渡される。

以下では、ブラウザ等から検索すべき文字列を LSE が受け取った後、検索が行われる様子を見ていく。LSE_1 と LSE_2、LS がある環境を想定し、LSE_1 がユーザーから検索要求を受け取ったものとする。

1. LSE_1 はユーザーの検索要求を受け取る。受け取った検索論理式は内部で最適化する。
2. LSE_1 は LS に検索すべき文字列を送信する。
3. LS は位置情報を検索する。
4. LS は得られた位置情報のうちスコアが高いものをいくつか返送する。ここでは説明のために LSE_2 だけを返送するものとする。
5. LSE_1 は LS から送り返された位置情報をもとに LSE_2 に検索を依頼する。
6. LSE_2 では LSE_2 が持っているデータを既存のサーチエンジンを使って検索する。
7. LSE_2 は検索結果を LSE_1 に送り返す。
8. LSE_1 は LSE_2 の検索結果を取りまとめ、検索結果をブラウザなどに送り返す。

CSE では LS があるキーワードに関してどこを捜せばよいかという位置情報を持つため、位置情報を定期的に更新する必要がある。また、CSE 全体として見たときに、インデックスの更新は一度に行えたほうが都合が良い。そこで、位置情報の更新に合わせて各 LSE が持つ文書のインデックスも更新する仕組みを用意した。

CSE の更新時の動作の概略図を Fig.2 に示す。

位置情報更新時には、LS が主体となって動作する。これは全体を制御するのに都合が良いからである。新しい LSE を LS に追加登録することを主な目的として、LSE が主体となって位置情報を更新することもできる。以下は更新時の動作の概略である。

†Nobuyoshi SATO, Takashi YAMAMOTO, Yoshihiro NISHIDA, Minoru UEHARA, Hideki MORI {jju, yama,nishida}@ds.cs.toyo.ac.jp, {uehara,mori}@cs.toyo.ac.jp
‡Dept. of Information and Computer Science, Toyo Univ.

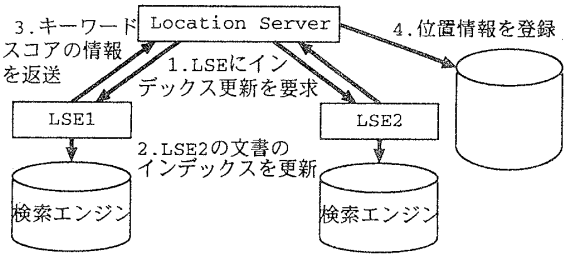

Fig. 2. CSE の位置情報更新時の動作

1. LS は各 LSE にインデックスの更新とキーワード、スコア情報の返送を要求する。
2. 各 LSE はインデックスを更新し、キーワードとスコアに関する情報を抽出する。
3. 各 LSE はインデックスの更新が終ったらキーワードとスコアの情報を LS に返送する。
4. LS はキーワード、スコアを位置情報として登録し、LSE の問い合わせに答えられるようにする。

4 異種サーチエンジンのスコア調整

CSE では現在の所、Namazu と SGSE の 2 種類のサーチエンジンを使っている。Namazu と SGSE ではスコアの計算方法が違うので本来はスコアの調整が必要である。Fig.3 に示すように Namazu のスコアに対する SGSE のスコアの比を計算すると、Namazu のスコアが 20 ポイント付近で小さくなったり、スコア毎の変動は大きいもののほぼ 1 前後である。また、各文書に対して双方のサーチエンジンが付けた文書に対するキーワード毎のスコアの違いは、概ね 1 から 3 程度の差で済んでいる。実際の検索ではこのスコアの違いはほとんど影響しないと考えて、現在の CSE ではスコアの調整は行っていない。しかし、将来的に Freya 等、他のサーチエンジンに対応したり、CSE 専用のサーチエンジンを作成する場合には調整が必要になると思われる。

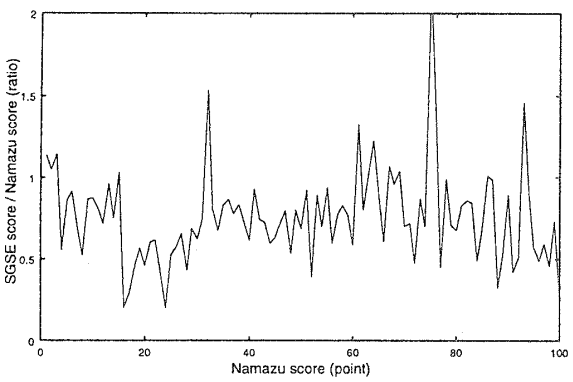

Fig. 3. Namazu と SGSE のスコアの比
Namazu のスコアが同じキーワード・文書に対して SGSE が付けたスコアを Namazu のスコア毎に平均した値の比

5 CSE における分散 tf·idf

tf·idf 法はキーワードが文書中に出現する頻度だけでなく、キーワードの『珍しさ』も考慮するスコアの計算方法である。tf はキーワードの単純な出現回数、idf はキーワードが一部の文書に集中している度合である。tf·idf 法を用いる際のスコアの計算式を以下に示す。

$$score = tf \cdot idf \quad (1)$$

$$idf = \log \frac{N}{n} \quad (2)$$

ただし、N は全文書数、n はヒットした文書数である。CSE において分散 tf·idf を用いる際の idf_{dist} の計算は以下のようになる。

$$idf_{dist} = \log \frac{\sum N_i}{\sum n_i} \quad (3)$$

ただし、N_i は各サーチエンジンの持つ文書数、n_i は各サーチエンジンでヒットした文書数である。

分散 tf·idf は既に WHERE[4] 等で用いられている。CSE では、idf の計算用に各 LSE が検索の結果返送する N_i、n_i を、tf の値として各文書のスコアを用いてユーザーに返送するスコアを計算している。tf·idf 法によるスコアの計算は最初にユーザーからの検索要求を受け取った LSE user interface がまとめて行う。このため、LSE 間の通信に用いるスコアはキーワードの出現頻度のみを考慮した値を用いる。Namazu はデフォルトの状態では tf·idf 法が既に適用されたスコアを出力するので、tf·idf 法適用前のスコアを出力するように設定して使用する。SGSE が出力するスコアは tf·idf 法を使っていないので、そのまま使用している。

6 まとめ

本稿では協調して検索を行う際のスコアの扱いについて述べた。今後、対応するサーチエンジンを増やす等の課題がある。

参考文献

[1] 馬場 肇『日本語全文検索エンジンソフトウェアのリスト』 http://www.kusastro.kyoto-u.ac.jp/~baba/wais/other-system.html

[2] 山名早人『*Trends of WWW Search Engines*』 http://www.etl.go.jp/~yamana/Research/WWW/survey.html (1998)

[3] 『次世代分散型情報検索システムに関する調査研究報告書』 http://www.jeida.or.jp/committee/jisedai/top.html (1997)

[4] Miguel Rio, Joaquim Macedo, and Vasco Freitas *"A Distributed Weighted Centroid-based Indexing System"*

[5] 高林 哲 『全文検索システム *Namazu*』 http://openlab.ring.gr.jp/Nnamazu/

[6] 『*SonyDrive Search Engine*』 http://www.sony.co.jp/sd/Search/SGSE-DL.html

[7] 山本崇、佐藤永欣、西田喜裕、上原稔、森秀樹『協調サーチエンジンの研究』,DICOMO'99,p169-174(1999)

[8] 西田喜裕、山本崇、佐藤永欣、上原稔、森秀樹『分散サーチエンジンにおける協調型検索』, SWoPP'99(掲載予定)