6X-6
~~6T-6~~

協調サーチエンジンにおける検索式の最適化

佐藤永欣 山本崇 西田喜裕 上原稔 酒井義文 森秀樹†
東洋大学工学部情報工学科‡

1 はじめに

ポータルサイトのような集中型検索エンジンでは、更新期間を十分短縮することは困難である。そこで、我々は分散型サーチエンジン Cooperative Search Engine (CSE) を開発した [1]。CSE では各 Web サイトに配された局所サーチエンジンが互いに協調して全体で一つのサーチエンジンを構成する。更新時にボトムアップで文書収集・インデックス作成が行えるため、更新時間を大幅に短縮する事ができる。しかし、検索時に余分な通信コストがかかる点が問題であった。この通信コストを削減するために検索対象サイトを最少化する手法を提案した [2]。しかし、まだ検索式自体は十分最適化されていなかった。そこで、本論文では CSE における検索式の最適化について述べる。

2 協調サーチエンジン

CSE は Fig.1 に示すように 4 つの要素で構成される。

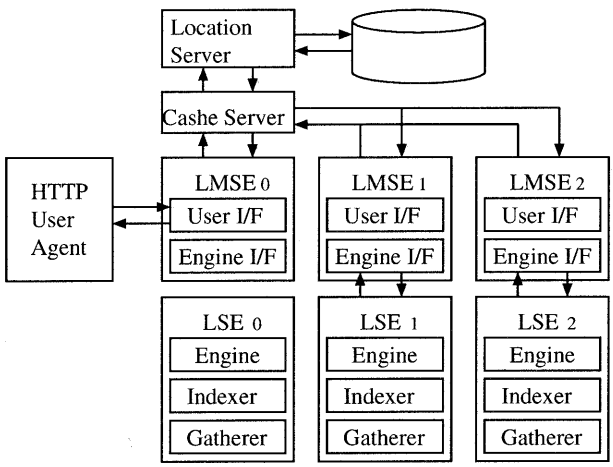

Fig. 1. CSE の概要

- Location Server (LS) は各 Web サイト野全文書に含まれるキーワードの表を管理するサーバである。
- Cache Server (CS) は検索結果をキャッシュするサーバである。
- Local Meta Search Engine (LMSE) は CSE のユーザー I/F と各種 LSE を隠蔽するエンジン I/F で構成された CGI である。
- Local Search Engine (LSE) は実際に収集、インデックス作成、検索を行う。CSE は Namazu と SGSE を使用する。

現在の CSE では以下のように検索を行っている。

1. ユーザーは LMSE_0 に検索式 E を送る。
2. LMSE_0 は E を CS に送る。
3. CS はキャッシュに検索結果があるか調べ、あれば返答し、無ければ LS に E を転送する。
4. LS は E 中のキーワードを含むサイトを検索し、検索対象のサイト集合 S を返送する。
5. CS は $\forall \mathrm{LMSE}_i$ in S に検索式 E を送る。
6. LMSE_i は検索式 E を LSE_i の検索式 E' に変換し、LSE_i に検索を依頼する。
7. LMSE_i は検索結果を整形して CS に返す。
8. CS は $\forall \mathrm{LMSE}_i$ in S の検索結果をマージソートして LMSE_0 に返送する。
9. LMSE_0 は HTML 形式に変換して返送する。

このように、検索式は最適化されていなかった。

3 検索式の最適化

CSE では検索対象を 2 段階で絞り込む。第 1 段階は、LS による検索式 E を送信する検索対象のサイト集合 S を決定する際の LMSE の絞り込みである。第 2 段階は、LMSE による文書の絞り込みである。第 1 段階の絞り込みについては文献 [2] で議論している。ここでは、第 2 段階の絞り込みにおける検索式の最適化手法について述べる。

前説で述べた通り、現在の CSE では検索式の最適化を行っていない。そのため、意味的に同じ検索要求が形式上異なる検索要求として解釈され、キャッシュサーバのヒット率を低下させていた。また、LSE での検索にも時間がかかっていた。

これらの問題を解決するため、CSE では検索式を 2 回変形する。第 1 の変形は、LMSE ユーザー I/F においてキャッシュサーバに検索要求を送信する前に行われる。この変形はキャッシュサーバのヒット率向上が目

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

的である。また、この式はそのままLSでの検索対象サイト絞り込みに用いられるため、サイト絞り込みに最適化されていなければならない。ただし、実装上はもう一つのバリエーションが考えられる。それは、キャッシュサーバが無変形の検索式を受信後、キャッシュする前に内部で変形する方式である。この方式ではサーバの情報を利用した最適化が可能となる。本論文では、後者の実装を採用する。

第2の変形は、キャッシュサーバにおいて対象となるLMSEに検索要求を送信する前に行われる。この変形は各LMSEの検索時間短縮が目的であり、LMSE毎に異なる可能性がある。

3.1 サイト絞り込みのための検索式最適化

検索対象サイト絞り込みは以下のように行われる。検索式をA、Bとすると、OR検索時にはA、Bを持つ全てのサイトに検索要求を送る必要がある。AND検索時にはAとBの両方を持つサイトにのみ検索要求を送れば良い。また、NOT検索時には、Aを持つサイトのみに検索要求を送り、Bををを持つがAを持たないサイトに検索要求を送る必要はない。

これらは検索式により形式的に回答可能なサイトを限定できる事を意味する。絞り込みのアルゴリズムは以下の通りである。

1. 積和標準形に変換する。
2. 矛盾のある項を削除する。
3. 各項内でキーワードを*idf*の降順に並び変える。ただし、NOTの右項は先頭に来ない。
4. 各項の*idf*を各項内のキーワードのうち、最小の*idf*で代表し、この降順に並び変える。

3.2 検索効率化のための検索式最適化

はじめに、検索確率の概念を導入する。検索式Eの検索確率$P(E)$とは、LMSEが検索式Eにマッチする文書を発見する確率をいう。キーワードk、検索式A、Bにおいて、

$$\begin{aligned} P(k) &= \frac{n_k}{N} \\ P(A \text{ AND } B) &= P(A) \times P_A(B) \\ P(A \text{ NOT } B) &= P(A) - P(A \text{ AND } B) \\ P(A \text{ OR } B) &= P(A) + P(B) - P(A \text{ AND } B) \end{aligned}$$

となる。ここで、n_kはkを含む文書数、Nは全文書数、idf(k)はkの珍しさ(inverse document frequency)である。CSEではスコア計算に$tf \cdot idf$法を用いているため、idf(k)は既知である。また、$P_A(B)$はAを含む文書内のBを含む文書の確率を表す。$P_a(b)$を求めるには二つのキーワードa、b間の相関を求めなければならないため、インデックスの情報量が増大する。また、検索式ではORよりAND、NOTの方が高い優先順位を与えられ、同じ順位同士では左結合性（最左演算子から評価される）があることが多い。そのため、ANDおよびNOTのみを含む項では最左キーワードの検索確率が最小であることが望ましい。そこで、項内の最左項を$P(k)$の最小のものとすることが有効な戦略となる。

検索式最適化のアルゴリズムを以下に示す。

1. 検索式を積和標準形に変形する
2. 検索対象のサイトLMSE$_i$ in S毎に以下の処理を行う。
   (a) 標準形の各項について以下の処理を行う。
      i. 項内のキーワードを$P(k)$の昇順にソートする。
      ii. 最左キーワードkの$P(k) = 0$となる項を削除する。
      iii. LMSE$_i$にその項を送信する。

4 評価

本方式の有効性を確認するためログ解析の予備調査を行ったところ、全体の6%が改善可能であった。また、語やシステムに依存するものの、頻度が1000、100、10の語を昇順と降順に並べた検索質問では、昇順の方が倍以上早い。

5 まとめ

本稿では、CSEにおける検索式の最適化手法について述べた。検索対象サイトの絞り込みは今までも行っていたが、各サイト内での検索の効率化のための検索式の最適化は行っていなかった。これにより、検索時間の短縮が可能である。

参考文献

[1] 山本 崇、佐藤永欣、西田喜裕、上原 稔、森 秀樹「協調検索エンジンの研究」DICOMO'99, pp.169-174(1999)

[2] 上原 稔、山本 崇、佐藤永欣、西田喜裕、森 秀樹「協調検索エンジンにおけるクエリー対象の最少化」DPSWS'99, pp.85-90(1999)

[3] 西田喜裕、山本崇、佐藤永欣、上原稔、森秀樹「キャッシュを用いた協調サーチエンジンの高速化」DICOMO 2000, pp.313-318 (2000)

[4] 佐藤永欣、山本 崇、西田喜裕、上原 稔、森 秀樹「協調サーチエンジンにおける継続検索のための先読みキャッシュ方式」 DPSWS 2000, pp.205-210 (2000)