協調検索エンジンにおける分散型スコアリング

上原、山本、佐藤、西田、森
東洋大学工学部情報工学科

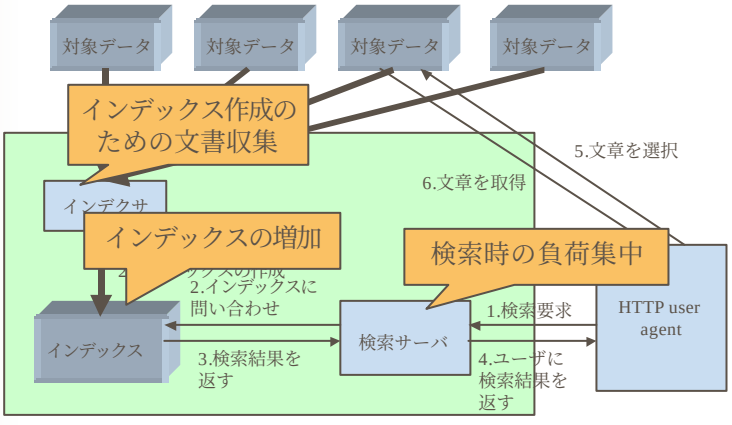
目次

- 背景及び目的
- 協調検索エンジンの概要
- 分散型スコアリング
- 実装
- 評価
- まとめ

背景および目的

- 組織内での情報検索
  - 更新期間の短縮→分散型検索へ
- 協調検索エンジンの提案
  - 検索エンジンによりスコアが異なる
  - 同種の検索エンジンでもtf*idf計算は正しく行えない
- 分散型tf*idfスコアリングが必要

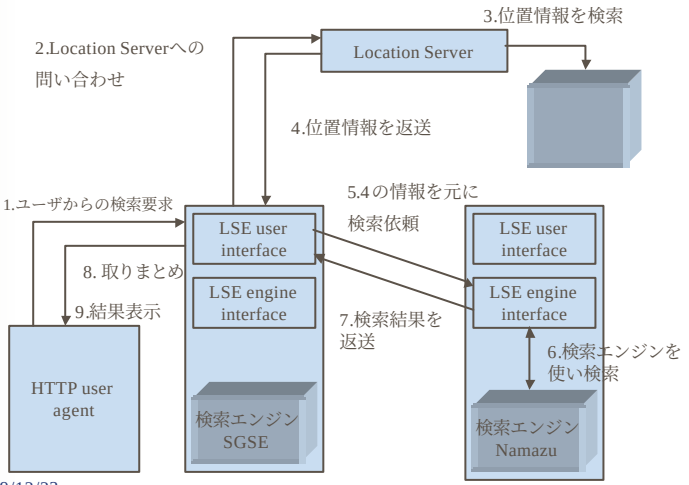
集中型全文検索システム



協調検索エンジンの提案

- 概要
  - 複数の検索エンジンを協調させ、１つの検索エンジンのように見せる
- 目的
  - 組織内情報検索の統一
  - 既存検索エンジンとの互換性
  - 更新期間の短縮
  - 集中型検索エンジンに劣らない性能

協調サーチエンジン



協調検索エンジンの問題点

- 基本検索エンジン（Namazu、SGSE）によりスコアが異なる
  - スコアの共通化が必要
- 同種のエンジンでもtf*idfスコアリングが正しく行えない

- 分散型tf*idfスコアリングの実現

tf・idf法

- s(d,k) = K × tf(d,k) × idf(k)
  - s(d,k)：キーワードkに対する文書dのスコア
  - K：タグなどによる重み付け
  - tf(d,k)：文書dにおけるキーワードkの出現頻度
  - idf(k) = log (N / n)
    - N：総文書数
    - n：キーワードkを含む文書数

キーワードの出現頻度だけでなく、珍しさも考慮したスコアの計算法

分散tf・idf法

- $s(d,k) = K \times tf(d,k) \times idf(k)$
  - s(d,k)：キーワードkに対する文書dのスコア
  - K：タグなどによる重み付け
  - tf(d,k)：文書dにおけるキーワードkの出現頻度
  - $idf(k) = \log (\Sigma N_i / \Sigma n_i)$
    - N_i：サイトiの総文書数
    - n_i：サイトiのキーワードkを含む文書数

N_i, n_iを集計する必要がある

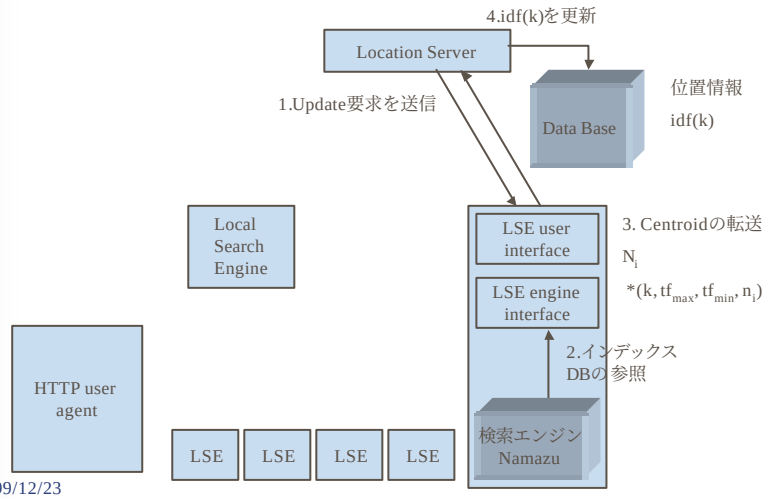
NamazuとSGSEにおける重み付けの相違

タグの種類	Namazu の重み	SGSE の重み
Keywords	32	100
description	32	10
title	16	2
H1~6	8~3	1
A	4	1
STRONG,EM,KBD,SAMP,VAR,CODE,CITE,ABBR,ACRONYM,	2	1
その他	1	1

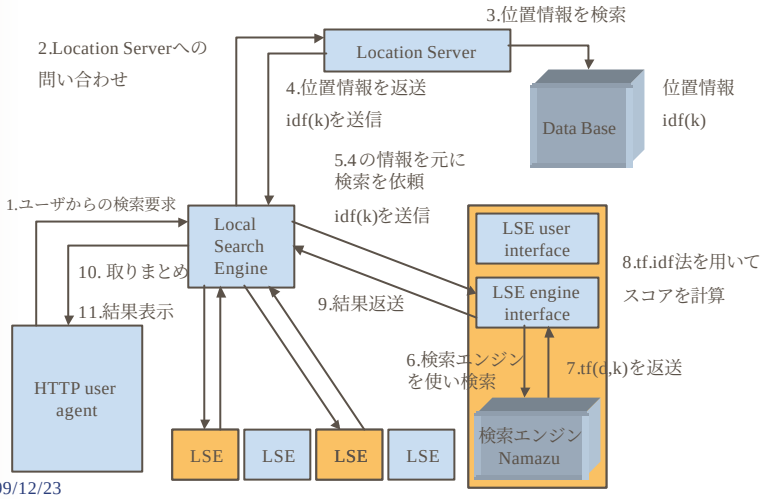
NamazuとSGSEにおける論理型検索におけるスコア計算

- Namazu
  - s(d,k) = K * tf(d,k) * idf(k)
  - s(d,A and B) = min(s(d,A), s(d,B))
  - s(d,A or B) = max(s(d,A), s(d,B))
  - s(d,A not B) = s(d,A)
- SGSE
  - s(d,k) = K * tf(d,k)
  - s(d,A and B) = s(d,A) * s(d,B)
  - s(d,A or B) = s(d,A) + s(d,B)
  - s(d,A not B) = s(d,A)

CSEにおける分散tf・idf法(1)



CSEにおける分散tf・idf法(2)

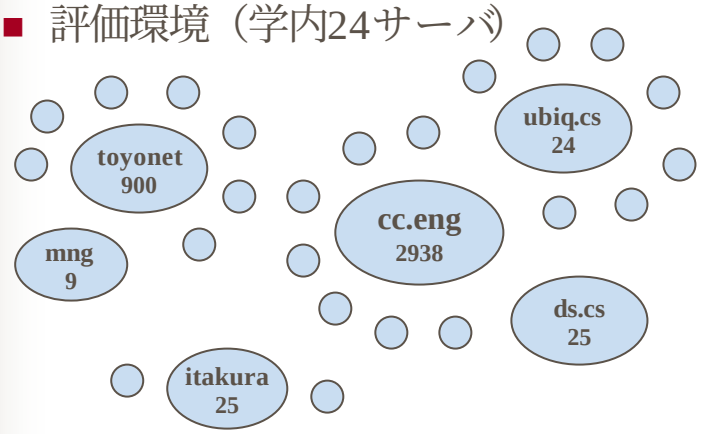


実装

- 方針
  - スコアリングの基準をNamazuに合わせる
  - インデックス情報のみを変換し、検索エンジン自体は変更しない（互換性・相互運用性）
- 方法
  - メタエンジン(LSE)が検索論理式を単独キー検索に変換し、論理型検索をエミュレーションする

実際の環境における評価(1)

- 評価環境（学内24サーバ）



実際の環境における評価(2)

	応答時間（秒）
CSE	1.267
Namazu	0.046
SGSE	0.448

単独サイトにおける単一キーワードによる最短検索時間の比較

		応答時間（秒）
AND検索	CSE	1.5
	Namazu	5.6
OR検索	CSE	5
	Namazu	5.6

CSEの並列検索とNamazuを用いた人手による逐次検索との比較

まとめ

- tf*idfに基づくスコアを分散計算
- 既存検索エンジンとの互換性・相互運用性を確保
- 通信遅延は大きいが、サイト集合の絞り込みにより実用的な性能を達成
- インデックスの更新時間は大幅に短縮

今後の課題

- 更新時間の更なる短縮
  - 高速なインデクサーの開発
- キャッシュサーバ
  - 次の１０項目を素早く返す
  - 規模の拡大（組織内から組織間へ）
- 分散型収集ロボット