Web アーカイブのアクセスログを用いたセマンティック検索アルゴリズム

2004MT083 濟田 将行 2004MT099 鈴木 淳雄
指導教員 河野 浩之

1. はじめに

Web アーカイブに保存されたデータで，諸問題を解決する論文は多く発表されている．しかし，実際の検索段階で，入力したキーワードに対して，全く違う内容の検索結果を表示してしまうという問題があるのが現状である．近年ではオントロジーと言う概念が注目されるようになってきたが未だ解決されていない．そこで我々は検索時の語彙の違いを解決し，検索結果をユーザの期待に沿うことが出来るような検索システムが出来ればユーザビリティの向上につながると考えている．

そこで本研究では現在アーカイブでの検索はどのように検索されているのか，ユーザの期待する検索結果をどれ程満たしているかを確認し，入力するキーワードに対しての関連語とカテゴリを用いた検索プログラムを実装し，検索結果の精度は適合率を用いて調べ評価する．

2. Web アーカイブとセマンティック検索

2.1. Web アーカイブ

Web アーカイブとは Web 上の情報資源を記録化し，その情報の内容と存在を空間的，時間的に安定化させるためのものである[1]. Web アーカイブは世界中で研究，実用化されている．その例として Internet Archive[*1] と WARP[*2] ついて述べる．

アメリカの Internet Archive は 1996 年からバルク収集で行われており，約 150TB という規模は世界最大級である．しかしバルク収集を用いているため，Web アーカイブには参照しても検索するユーザにとって期待通りに結果を出力しない事が多々ある．それは時系列がきちんと整理されていない事や，ページ更新等が管理者によってタイミングがバラバラなことに原因がある．それに対して WARP では選択的収集を行っており，現在約 19TB の規模である．

2.2. Wera[*3]

Wera（Web Archive Access）とは Web アーカイブの管理ツールである. Internet Archive の Wayback Machine のように Web アーカイブでの検索の際に更新前のページも閲覧可能となっており，利便性は高くなっている．そして，Wayback Machine のように URL を入力する必要がなく，キーワードの検索が可能である．また，時系列ページの閲覧の際に時間の指定の必要がなく，好きな時間帯を選ぶことができるため時系列検索が容易であるといった点がある．しかし，Wera にはセマンティック検索ができておらず，また関連語検索が不十分であるといった問題が存在している．

2.3. セマンティック検索

既存の検索システムはユーザのキーワードを入力し，検索をクリックして検索結果から情報を得ている．しかし，それだけの検索システムではユーザの意思の沿った検索結果を出力することは難しい．例えば「ATM」について検索をした場合，「ATM」は「Automatic Teller Machine」のことなのか，もしくは「Adobe Type Manager」なのか判断できない．よってユーザが何を考えているかを検索システムに考慮させる必要がある．それを実現する方法として，現在オントロジーを用いた検索システムが考えられている．その 1 つに SW-IQS が提案されている[2]. これは自動分類法とランキングアルゴリズムを用いてユーザの期待に沿う結果を出力している．

また現在の検索エンジンにおいてのセマンティック検索ではカテゴリ検索がある．ユーザのキーワードに対し，木構造になっているカテゴリの中からそのキーワードを検索していく．これはユーザの知りたい情報に対して，カテゴリを掘り下げていくことによってキーワードの意味を確定し，検索結果を出力している．

2.4. Namazu

本稿では日本語全文検索システムである Namazu の概要を説明する．cgi として動作させることにより，小中規模の WWW 全文検索システムを構築することができるほか，ハードディスク内のファイルを対象としたパーソナルな用途にも使えるようになっている UNIX 系のフリーソフトである．Namazu はインデックスという索引ファイルを用いているため高速な検索が可能となっており，なおかつ検索結果を簡単にカスタマイズできるため，自分の好きな用途の検索システムの作成が可能となっている．また，語句の重み付けには TF-IDF 法を用いており，文章の語句検索においてユーザのニーズに応えた結果が表示される．

Namazu はインデックス作成に mknmz コマンドを用いる．mknmz を実行したとき「NMZ.*」ファイルがたくさんできるが，このファイルのひとまとまりが 1 つのインデックスとなっている．本研究では namazu.cgi をカスタマイズするのではなく，Namazu 関数を用いて，検索プログラムを作成する．

[*1] Internet Archive, http://www.archive.org/index.php

[*2] WARP, http://warp.ndl.go.jp/

[*3] Wera – Homepage, http://archive-access.sourceforge.net/projects/wera/

3. 検索アルゴリズムの提案

3.1. セマンティック検索アルゴリズム

現在 Web アーカイブはセマンティックにおける問題がある．その問題を解決するために，先行研究[2]で提案している SW-IQS は文書ページの検索にセマンティックな役割を持たせている．また，先行研究[3]ではアクセスログを用いることでキーワードにおける関連語を発見している．

本研究では先行研究[2][3]で提案されているアルゴリズムを組み合わせ，改良した新たなアルゴリズムを提案する．アルゴリズムにおけるフローチャートを図 2 に示し，以下にその説明を行う．

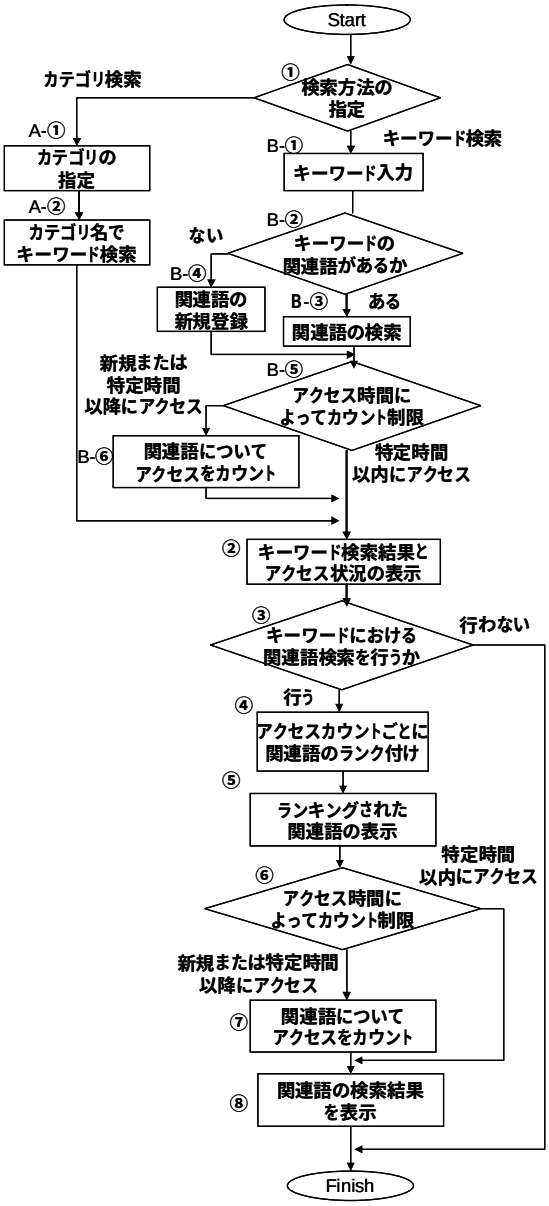

図 2　検索アルゴリズムのフローチャート

例として「温泉」という語句でキーワード検索したときの動きを図 2 を用いて説明する．まず①で「温泉」に対してキーワード検索かカテゴリ検索か判定をする．カテゴリ検索を行う場合，A-①で「温泉」に対するカテゴリを指定する．例えば「地域別」，「温泉ガイド」といったカテゴリが検索される．そして A-②で「温泉 → 温泉ガイド → 愛知県」というようにカテゴリ名を決定し，そのカテゴリ名でキーワード検索を行う．「温泉」というキーワードに関して関連語検索を行う場合は B-①でキーワードの入力を行い，B-②でキーワードに対しての関連語が存在するかの判定を行う．例としては「温泉」からは「スキー」，「民宿」などといった関連語が B-③で検索される．しかし，関連語が存在しない場合は，B-④の処理を行う．関連語の検索の後，B-⑤でキーワード検索におけるアクセス時間の判定を行う．そしてアクセスが特定時間以内のときは，何も処理をせず，“新規アクセス”，または特定時間以降にアクセスを行った場合は B-⑥でアクセス数のカウントを行う．先ほどの例では，「露天風呂」という関連語に対し，アクセス数を加算する．その処理後，②で検索結果を出力する．さらに③で関連語検索を行うか判定をする．そして検索を行う場合は④でアクセス数ごとに関連語のランク付けを行う．「温泉」では「スキー」「民宿」といった関連語が，一般的にユーザが関心を持つ関連語で上位にランキングが行われる．そして⑤でそれらの関連語を表示し，⑥で関連語検索についてのアクセス時間の判定を行い，特定条件の場合⑦でアクセスのカウントを行う．ここの⑥，⑦の処理は B-⑤，B-⑥の処理とほぼ同様である．そして⑧で検索結果を表示して終了する．

4. アーカイブ検索プログラムの実装

我々は提案したアルゴリズムを基に PHP によってプログラムを作成し，アーカイブコレクションを想定した実行環境で実装し，評価を行う．本稿ではプログラムの実行環境について述べる．

4.1. アーカイブ検索プログラムの実行環境

本稿ではキーワード検索プログラムの概要を述べる．我々は index.php, result.php の 2 つのプログラムを作成した．

(ア)　index.php

検索語句を入力し，その情報を検索ページへ送るプログラムである．検索がキーワード検索の場合は情報を result.php へ，カテゴリ検索の際は semantic.php へそれぞれ情報を送る．

(イ)　result.php

index.php で送られてきたキーワードを基に，Namazu を用いて検索し，検索結果を表示するプログラムである．複数の言葉で検索したときは，最初の語句をキーワードとし，次の語句をその関連語とする．そしてキーワードとその関連語がデータベースに登録されているかの確認

を行う．登録されていなければその情報をデータベースに登録する．そして検索の際にはNamazuを用いて検索を行う．そしてページ情報を取得し，結果を表示させる．

図3　index.php で作成したトップページ

ここの例では「Web 検索」での検索を行っている．図3 でのトップページで「Web 検索」でキーワード検索し，検索結果を図 4 で表示させている．そして検索結果画面にはユーザが新規アクセスか特定時間内のアクセスのアクセス状況とキーワード検索画面，さらに「Web」，「検索」という語句における関連語の検索が行うことができるようになっている．

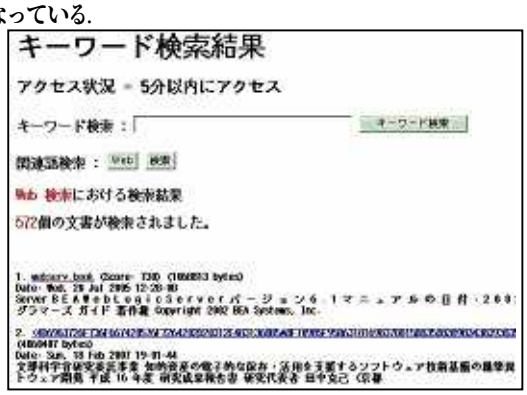

図4　result.php で作成したキーワード検索結果画面

4.2. 関連語検索プログラム

本稿では関連語検索の詳細を実際に紹介する．我々は relation-result.php という関連語検索プログラムについて述べる．これはキーワードの関連語を表示するプログラムである．関連語 DB へキーワードを送り，キーワードに対した関連語が存在するか判定を行う．そして存在した場合はその関連語のタグをアクセス数上位 10 位毎に表示し，存在しない場合は関連語が存在しないことを示し終了する．関連語タグをクリックすると，キーワードとその関連語を log.php へ送り，ジャンプする．尚 log.php については 4.3 節で紹介する．

図5　relation-result.php での関連語検索画面

図5では図4でのキーワード検索結果から「Web」で関連語検索したときの画面である．ユーザが複数でキーワード検索，または関連語検索したときのアクセスカウントごとにランキングされており，この図より現在多くのユーザが関心をもっている関連語が「情報」であることがわかる．

4.3. アクセスログ制限プログラム

アクセスカウントでランク付けを行う検索では，連続アクセスなどの悪質な荒らしがあった場合，正確なランキングを表示することができない．そのため，関連語検索におけるアクセスログ制限のプログラムを紹介する．我々はアクセス制限に関連語検索で用いた relation.php を用い，またアクセスログ制限を行うための log.php を作成した．log.php は relation-result.php から送られてきた関連語に IP アドレスと時間によってアクセスカウントの制限を行う．IP アドレスが未登録なら新規にアクセスログＤＢに登録し，新規アクセスと返す．既に登録されていた場合，時間で判断する．5分経ってからアクセスをしたらユーザのアクセス情報である時間と IP アドレスの更新を行い，カウントを加算する．しかし，5 分以内にアクセスをした場合はアクセス情報の更新は行わず，カウントの加算は行わない．そのため，このシステムでは検索時に連続アクセスなどの悪質な荒らしを避けることができた．

4.4. カテゴリ検索プログラム

我々はカテゴリ検索を行うため，Namazu で使用されているインデックスを複数作成し，それらを第 1 カテゴリとして検索を行う．また，カテゴリは 3 階層で作成をし，カテゴリ階層については図 6 に示す．

例えば「検索」という語句で検索を行うとき，「Web アーカイブ」というカテゴリで検索すると，検索結果として「オントロジーを用いた検索」などが表示される．しかし，「P2P」で検索を行うと「ノード検索」などが表示される．我々はプログラムとして semantic.php を作成した．これは index.php で決定した第 1 カテゴリを基に，各データベースから第 2 カテゴリの語句を表示する．また，第 2 カテゴリのキーワードで検索を行う場合，再帰的に semantic.php で検索し，各デー

タベースから第 3 カテゴリの語句を表示する．カテゴリでのセマンティック検索を終了するときはキーワードと各カテゴリの語句について AND 検索を行い，結果を出力する．

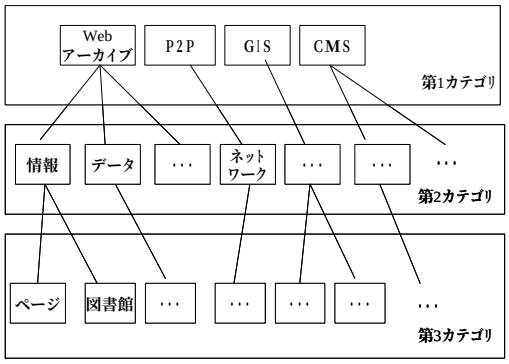

図 6　カテゴリの階層

5. アーカイブ検索プログラムの評価

本稿ではカテゴリ検索プログラムの検索結果を基にアンケートを行い，その結果から情報検索システムの評価方法としてよく用いられる適合率を求め評価する．適合率とは検索結果の中の適合文章数を検索結果の文書数で割ったもので，この適合率は検索結果がユーザの質問，考えている事に合っているかどうかを算出していることから，我々の実装したセマンティック検索システムの評価に適している．

関連語検索システムのアンケート方法は，回答者に対して「CMS」と「Web アーカイブ」というキーワードに対してそれぞれの関連語に興味が有るか，もしくは無いかを選択をさせた．アンケートの詳細は以下のようになる．

- ･対象　南山大学数理情報学部情報通信学科に所属する学生 10 名
- ･目的　実装した検索システムの結果がユーザの意図に沿っているかを評価
- ･方法　調査は質問用紙により実施
- ･内容　① 関連語検索はキーワードを入力して，表示された関連語に興味があるかないかを調査
  ② カテゴリ検索は検索を実際に行い，結果をみて実際に調べようとしていた内容に合っているかどうかを判断

アンケートを行った結果，得られた数値を表 1 に示す．

検索システムを評価する適合率は検索結果として得られた集合中にどれだけ検索に適合した文書を含んでいるかという正確性の指標である．表 1 より我々はセマンティック検索システムを実装し，全体で適合率 70.5%という数値を得ることが出来た．関連語検索システムでは適合率が 75%となった．これは収集したデータを形態素解析し出現率の多い名詞を抽出することによって得られたものが，検索をするユーザにとって興味のある語句となり，ユーザの期待に沿える検索システムになっていると考えられる．一方カテゴリ検索システムでも 66%と約 3 つに 2 つは期待通りの検索結果が出力していると評価することが出来た．

表 1　アンケート結果における各検索の適合率 (%)

	キーワード別 適合率	検索システム別 適合率
関連語検索 (Web アーカイブ)	73	66
関連語検索 (CMS)	77	
カテゴリ検索 (Web アーカイブ)	62.5	75
カテゴリ検索 (CMS)	69.5	
システム 全体の適合率	70.5	

6. まとめ

我々はWebアーカイブにおいてセマンティック検索を行うためにアクセスカウントを用いた関連語検索，語句の出現数からカテゴリを決定するカテゴリ検索の2種類を提案し，実装，評価を行った．その結果関連語検索とカテゴリ検索を組み合わせることで適合率は70.5%となり，セマンティック検索を行うことができた．

この結果から考察すると，ユーザのアクセスカウントのランキングと語句の出現数からユーザのニーズに沿ったセマンティック検索を行うことが可能となることがわかる．

今後の課題としてはオントロジーを文書の出現数とは別の方法で作成し，我々の提案したアルゴリズムと組み合わせることでセマンティックの精度がさらに上がるのではないかと思われる．また，さらに多数の文書ページを収集した場合，カテゴリが今の階層だけでは不十分であるためカテゴリの拡大とインデックスの更新が必要であるといえる．

参考文献

[1] 廣瀬信己，“国立国会図書館におけるウェブ･アーカイビングの実践と課題，” 情報処理学会研究報告，No.51, pp.95-111, 2003.

[2] Okkyung Choi, Sangyong Han, Ajith Abraham, “Integration of Semantic Data Using a Novel Web Based Information Query System,” International Journal of Web Services Practice, Vol.1, No.1-2, pp.21-29, 2005.

[3] 大塚真吾，喜連川優，“大規模アクセスログを用いた検索支援システム，” データ工学ワークショップ，DEWS2006, 1B-o2, 2006.