2H5-12

利用者のコンテキストに従ったコンテンツ検索・編纂システム

Content-Compilation System based on User's Contexts

嶋津恵子 Keiko Shimazu
齊藤功 Isao Saito
吉永早織 Saori Yoshinaga

慶應義塾大学デジタルメディア・コンテンツ統合研究機構
Research Institute for Digital Media and Content，Keio University

Nowadays，new demand for content-search engine on Internet is appearing，which is for rearrangement/realignment search-result content based on user's context. We designed this system consisted of 2 layers. One is intra-university network system and the other is context generations. The users of this system can look down at the search result content, know the variation on an information maps and notice the relationship among contents. The system with this function will contribute not only to user's request of acquiring information，but also user's request of supporting their idea-creation. We call this new function on content search system" Content-Compilation". In this experimental study，we focused 4 research issues for an evaluation test. We enquired information of remarkable places (cites and/or area) name for each issues. As a result，we confirmed our system's contribution for information acquisition task.

1. はじめに

インターネットの台頭と世界規模のブロードバンド化により，発信・流通される情報量が爆発的に増加している． 米カリフォルニア大学 バークレー校の School of Information Management and Systems の発表によると 2001 年まで世界中の人間が一人当たり 250 メガバイトのデータを作成し[1]，2002 年には全世界約 2 万 3,000 ペタバイトの情報が生成された[2]． これにより日常生活にも大きな変化が発生し，「知りたいことやわからないことはまず，インターネットで調べてみる」人は 92.4%に上るという[3].

その一方で，望む情報を効率的に探し出すことができない，つまり「必要とする情報が見つからない」「読み落としてしまう」といった情報洪水の指摘も生まれている． さらに，最近では必要な（どこかに存在するはずの）情報を探し出す機能の強化とは異なる，新たな要求が発生し始めている． これは特に人手によるサービス提供の領域での新ビジネスの台頭にみられる．例えば，サーチャはデータベースから対象情報を的確に探し出すことが主な役割であるが，最近ではこの発展形としてコンシェルジェが注目され，さまざまな分野で活躍している．

特に情報の有効活用という点ではブックコンシェルジェを注目すべきである[4]． ブックコンシェルジェは，顧客の要望から領域やキーワードを特定し書籍を検索するだけでなく，顧客の仕事の傾向や趣味趣向を反映させ（既存の分類項目や検索用の用語で特定できない）書籍を世界中の流通対象から選定する．さらに一部のブックコンシェルジェは，該当する書籍の購入だけでなく，顧客が既に所有している書籍類が納められている書架へ，これらの新規購入書籍を配置するまでを行う．このサービスの特長は，購入した書籍を空きスペースに格納する技術ではなく，すでに設置してある他の書籍との関係を，配置によって表現する技術の提供にある．

この付加価値提供が成功しているのは，顧客が所望している情報がもはや特定の書籍ではないという事情を受けている．彼らは，特定の書籍から情報を得るだけでなく，一定のコンテキストに従って配置された書籍群を俯瞰することで，新たな発想のきっかけを得ようとしている．

我々は同様の現象がインターネットのコンテンツ検索にも起こりつつあると考える．つまりキーワード指定による汎用的な検索により抽出さえたコンテンツを，コンテキストに従って俯瞰する最適な方法の提供が重要になってくる．これはリスト表示順位の入れ替えだけでなく，コンテキストを反映した情報マップとその上への配置手法の開発が必要になる． そこで我々は，高度化する検索アルゴリズムの応用技術の開発を引き続き目指す一方で，何らかの条件（現在は特定の共通の用語を持つ）で抽出されたコンテンツを，利用者のコンテキストに従って並び替え，再配置するという新しい機能の開発を目指している．我々は，これらを"コンテンツの検索・編纂"と位置付けている．これにより，コンテンツ検索システムは，利用者に新たな気づきを与えたり，価値あるコンテンツの効率的生成を支援するシステムに成長すると考える．

今回我々は，この第一歩となるシステムを試作した．そして，4つの話題を取り上げ，それぞれに関する注目すべき重要な場所（地名や都市）情報を獲得することを目的に実験をおこない，コンテンツの検索・編纂の効果を確認した．

本書は以下の構成をとる． 2 章でインターネットコンテンツ検索の課題とその解決を目指す先行事例とを記す． 3 章に開発したシステムの構成を述べ，4 章にその有効性の検証テストを示す． 5 章に考察を，6 章にまとめを述べる．

2. 汎用検索エンジンの課題と先行システム事例

Google の検索エンジンが万能でない顕著な例は，アジア，特に中国，韓国，日本における検索サイトの人気順位に示される． いずれも Google は一位となっておらず特に中国で百度が圧倒的な支持を受けている． 百度は音楽ファイル（MP3）検索や携帯向け検索に強い機能を提供し，またコミュニティ支援機能にも評価が高い． また，韓国で人気の検索サイトでは，検索キーワードを入力するとカテゴリーごとに結果が表示され，利用者は自身のコンテキストに合致したページに移動する． 従来

[1] http://www. Brocadejapan.com/news/pr040501.hp

[2] http://internet.watch.impress.cojp/cda/news/2003/11/04/990.html

[3] http://japan.cnet.com/news/media/story/0,2000056023,20242028,00.htm

[4] Ｌマガジン 12 月号，（株） 京阪神エルマガジン社,2006

連絡先：shimazu@dmc.keio.ac.jp

の汎用型の検索エンジンの最大の課題 である「検索結果には全く関係ない同音意義語が多く混在する」という問題に対し，検索結果のコンテンツを，いったんコンテキストで分類・構成するという方法で解決を図っている．

欧米でも，利用者のコンテキストを反映するという課題に注目したサービスの提供が始まっている． GeoFusion は検索で得られたデータを図式化して表示する機能を提供し，特に地理データを扱うことによる検索結果の有効利用を提案している． 例えば，タグを取りつけたマグロの動きを示す地図を描き出すことが可能である．

また，カリフォルニア大学バークレー校は，芸術作品や骨董品の検索を容易化した「Flamenco」と呼ばれる検索エンジンの試作版を発表した． 世界中の美術 館が所蔵するコレクションから集めたアート作品が，内容（たとえば，動物，天国と地上，形状や色など）や，製作年代，作者，メディア（絵画， 家具，彫刻など）の共通の属性に従って，検索結果が分類表示される．利用者は，コンテンツを異なる視点で並び変えて参照することが可能であり，例えば 1700 年から 1709 年の間に絞りコンテンツを参照すると，この期間には蹄をもつ動物の絵が 4 枚しか描かれていないことを発見する[1]．

さらに国内では，これらと同様の課題に注目した専用の検索サービスが試行開始された[2]． ここでもキーワードを用いて検索した結果を，まず一定の体系で分類し，その見出を利用者に提供する． これにより現在の検索エンジンによる普遍的な問題，つまり多すぎる出力件数や絞り込みに最適なキーワードが思いつかないことによる，知りたい事柄にたどり着けない問題の解決を狙っている．

学界からは，汎用の検索エンジンによる出力結果を，利用者のコンテキストに従って表示順位を並び替える手法の提案が多い[Hoeber 06][Yao 02]． これらはいずれも，リスト形式に出力された検索結果の順位を変え，別の基準で並び変え，同じリスト状に表示しなおすものである．

一方我々が研究目標としているのは，検索結果として出力されたコンテンツをコンテキストに最適な別の表示系に再配置する手法の開発である．

コンテキストに従って検索結果のコンテンツを並び替えることが有益であることは， 直観的には理解されているが，従来の検索エンジンによる結果と具体的に比較した報告は少ない． そこで今回我々は4つの話題を取り上げ，それぞれに関する注目すべき重要な場所（地名や都市）情報を獲得することを目的に，コンテンツの検索・編纂の効果を確認した．

3. 実験システム概観

我々は，コンテンツの検索・編纂システムを全学用コンテンツ流通システムのアプリケーションとして設計した． つまり，全学ネットワーク上のコンテンツをそれらが持つ用語をキーとして検索する基盤システム上に，検索したコンテンツをコンテキストに従って並べ変える機能を搭載した．

3.1 Intra-Univ. Content-Sharing System

我々はコンテンツ検索の基盤システムとして学術版のエンタープライズ・コンテンツマネジメントシステムを構築した(図 1 参照)． 具体的には，全学ネット上に情報収集ロボットを設置し，keio.ac.jp および keio.edu を巡回した(図 1 の(1))． 収集したドキュメントごとにテキストデータから固有名詞を抽出し，検索用インデックステーブルにタグ情報としてアノテート（付記）する(図 1 の(2))． コンテンツを検索したい利用者が，対象コンテンツが持つと予想される語(キーワード)を指定すると，システムは条件を満たすものを出力する(図 1 の(3))． このとき出力対象コンテンツの表示順は，システムがコンテンツの重要度を計算し決定するが，Google は独自のペイジ・ランク・アルゴリズムを導入したことで実用性の高さにおいて他を圧倒した[山名 05]． 一方我々のシステムは，インデックステーブルのタグ情報をコンテキストで分類しておくことで，利用者の目的に応じてコンテンツの順位決定や再配置を実現している．

次節に，研究分野にとって重要な場所や都市に注目したコンテンツの再配置機能を実現するモジュールを示す．

3.2 コンテンツの編纂

我々のシステムは，前節で述べたようにクローラーがネットワーク上の情報を収集した結果を対象に，固有名詞の抽出を行う．具体的には収集した情報がドキュメントごとに分類され，さらにそれらが持つテキストが形態素解析され 20 種類の名詞が特定される．コンテンツ検索時には，これらの組み合わせによりコンテキストを表現することが可能になる． 例えば，「"1990 年代"に"村井純"が関与する"インターネット"に関する情報」というコンテキストでコンテンツを指定することができる．

今回，我々はさらに慶應大学のコンテンツに特化したコンテキストを表現する機能を搭載した． 具体的には，研究上重要な地名や地域及び場所を”学術関連地名”[3]とし，これらをコンテンツが持つ場合，その緯度経度の計算結果とともに，検索インデックステーブルにタグとしてアノテートした(図 1 の(4))． 利用者が任意のキーワード（学術名や研究名称）でコンテンツを検索し，それらを地図上に再配置する操作をおこなうと，学術関連地名が存在するコンテンツだけを対象に，それらの緯度経度情報を算出し，GoogleMapsAPI を介し世界地図上に配置する(図 1 の(5))． 換言すると 我々の開発したコンテンツ編纂機能は，一旦キーワードで検索されたコンテンツ群に対し，指定したコンテキスト（今回は学術関連地名）で篩にかけ，該当するものを対象に専用の GUI（今回は世界地図）上に表示する． 利用者は，示された場所をクリックすることで対象のコンテンツを入手できるだけでなく，専用のUI上でコンテキストに従ってコンテンツ群を俯瞰することができる．たとえば本書で述べる実験では，検索キーワードで指定された話題に特定し，学術関連地名というコン

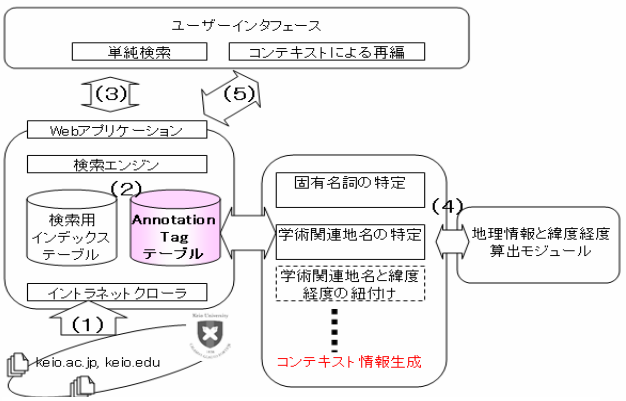

図 1. 全学コンテンツ検索・編纂システム概観

[1]http://japan.cnet.com/news/media/story/0,2000056023,20070366,00.htm

[2]http://www.hitachi-system.co.jp/press/2006/pr061031.html

[3]例えば，"ピラミッド"研究に対する"エジプト"がこれに相当する．

テキストで世界地図上にコンテンツ群を配置する. これにより, それぞれの話題にとって重要な場所が世界にどう分布/偏っているかを鳥瞰することができる.

図 2 に 5.1 節で述べたキーワード検索結果の出力例を示し, 図 3 に本節で述べたコンテキスト(今回は学術関連地名)でコンテンツを地図上に配置しなおした結果例を示す. 前者の状態では望む情報がどのコンテンツに記載されているかは, コンテンツを開き内容を確認しないと特定できない. 一方後者の手法を採ると, 一見してそれが世界中に点在し, かつアジアに重点が置かれていることが理解できる.

この機能の情報共有の有効性の実験を次章に述べる.

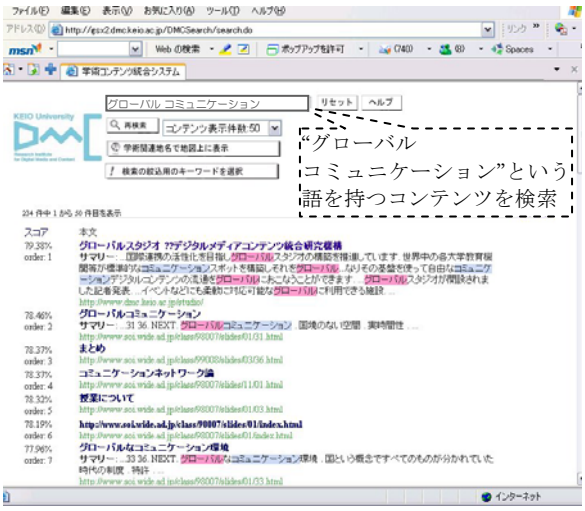

図 2. コンテンツの検索の出力画面例

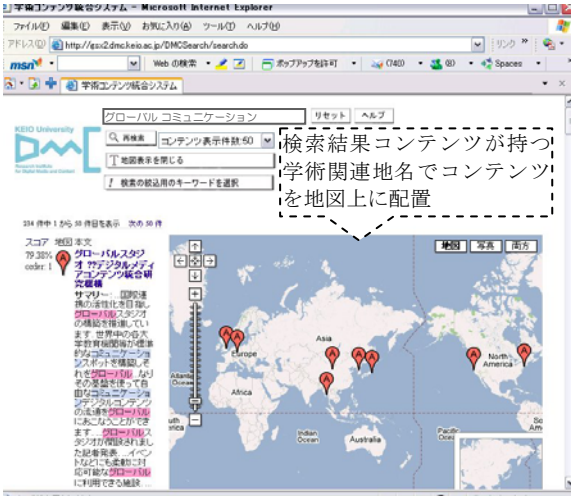

図 3. コンテンツの編纂の出力画面例

4. 有効性の検証

我々は, 今回開発したコンテキストに従って検索結果コンテンツを地図上に再配置する機能の有効性を検証するために, 次の 4 つのケースを用意した.

i. “環境計画”の学術関連地名を知りたい.
この情報そのものを説明したコンテンツは対象ネットワーク上に存在せず, 3 件のコンテンツを入手することで所望する情報(8 つの学術関連地名)を入手できる.

ii. “近代日本経済史”の学術関連地名を知りたい.
この情報そのものを説明したコンテンツは対象ネットワーク上に存在せず, 2 件のコンテンツを入手することで所望する情報(10 個の学術関連地名)を入手できる.

iii. “グローバルコミュニケーション”の学術関連地名を知りたい.
この情報そのものを説明したコンテンツは対象ネットワーク上に 1 つ存在し, このコンテンツを検索すれば所望する情報(8 つの学術関連地名)を入手できる.

iv. “地理学”の学術関連地名を知りたい.
この情報そのものを説明したコンテンツは対象ネットワーク上に 1 つ存在し, このコンテンツを検索すれば所望する情報(10 個の学術関連地名)を入手できる.

これら4つのケースを対象に, 従来のキーワードの追加による絞り込み検索と, 今回開発したコンテンツ編纂機能を, 入手したい情報を獲得までに要する操作数で比較した.

表に示すように, ケース i, ii, iv で必要な情報を獲得するのに, 我々が提案する機能を利用した方が, 従来のキーワードの追加による絞り込み検索より少ない数値(手順数)を示している. 3 ケースとも従来手法を用いると, 10 以上の手順(特に i に関しては 15 手順)を必要とするが, 我々の提案する機能を用いると 10 以下になる. 一方ケース iii は, 従来手法を用いた方が操作手順数は少ない.

5. 考察

5.1 入手したい情報が複数のコンテンツから構成される場合

多くの場合必要とする情報は, インターネット上に存在する. ただし, その情報が URI で特定できる一つのドキュメントに記載されている場合と, 複数の(URI から成る)ドキュメントに跨っている場合がある. 例えば, 地球温暖化に関しその問題の概要を知りたい場合, 多くのコンテンツが初心者向けの解説を述べており, そのうちの一つを検索できれば利用者は望む情報を獲得できる. 一方, 神奈川県藤沢市の市民生活の年ごとの変化と, 地球温暖化の関係を知りたい場合, これに即して論じられている単一のコンテンツが存在しないことも考えられる. その場合, 利用者は部品となる情報が記載されているものを集め, 必要に応じてそれらを組み合わせる.

今回の実験のケース i と ii がこれに想定する. つまり両ケースとも, それぞれの研究テーマにおいて重要だと判断される地名や場所に関して説明している単独のコンテンツは(慶應大学の)ネットワーク上には存在せず, この情報を所望する利用者は複数(ケース i は 3 件, ii は 2 件)のコンテンツを入手することで要求が満たされる.

このような場合に, 今回開発したシステムが効果を発揮したことが, 必要な情報にたどり着くまでに要した操作手順数の差からわかる. 両ケースとも必要な操作手順数が約 3 分の 1 となっている.

5.2 入手したい情報が単一のコンテンツに述べられている場合

ケース i と ii に対し, iii と iv は望まれる情報が 1 つの(URL で特定できる)ドキュメントに搭載されているが, iii と iv で異なる結果となった. 具体的には, iii が最短の手順(検索結果の先頭のドキュメントを開示)で望む情報を獲得できたのに対し, 後者は, 16 操作を要している. これは iv の場合, 最初のコンテンツ検索で先頭 10 件中に望む情報を持つコンテンツが出現しなか

表. コンテンツの編纂機能の効果の検証

	コンテンツ検索			入手したい情報にたどり着くまでの次の方法			必要操作数
	入手したい情報	検索用キーワード	出力件数				
i	近代経済史にとって重要な地名や都市名	経済 歴史	500	キーワードの追加	追加した検索 キーワードの 数	ドキュメント開示操作数	
					1	14	15
				コンテンツの編纂	専用GUI(地図UI)の拡大/縮小操作数		
					6		6
ii	環境計画にとって重要な地名や都市名	環境 計画	500	キーワードの追加	追加した検索 キーワードの 数	ドキュメント開示操作数	
					1	18	19
				コンテンツの編纂	専用GUI(地図UI)の拡大/縮小操作数		
					5		5
iii	グローバルコミュニケーションにとって重要な地名や都市名	グローバルコミュニケーション	234	キーワードの追加	追加した検索 キーワードの 数	ドキュメント開示操作数	
					0	2	2
				コンテンツの編纂	専用GUI(地図UI)の拡大/縮小操作数		
					10		10
iv	地理学にとって重要な地名や都市名	地理学	117	キーワードの追加	追加した検索 キーワードの 数	ドキュメント開示操作数	
					1	16	17
				コンテンツの編纂	専用GUI(地図UI)の拡大/縮小操作数		
					6		6

【 用語説明 】
キーワードの追加:
　キーワードの追加による検索の絞り込みをおこない,
　出力されたコンテンツを人手で内容確認しながら入手したい情報にたどりつく方法
コンテンツの編纂:
　キーワードの追加による検索の絞り込みは行わず,
　汎用のコンテンツ検索を行った結果に対し,
　学術関連地名を対象に地図上に配置する方法
追加した検索キーワード数:
　検索の絞り込みのための追加したキーワードの数
ドキュメント開示数:
　入手したい情報がコンテンツに含まれているかどうかを確認するために
　開示したドキュメント数
**上記2操作は,
　通常利用を想定し, 検索結果の出力数を10件/1ページとし,
　上から順にコンテンツを開示する.
　1ページに出力される最後のコンテンツ(10件目)まで開示しても
　入手したい情報が得られない場合, キーワードを追加し絞りこみ検索を行った.
専用GUI(地図UI)の拡大/縮小操作数:
　入手したい情報を得るために必要となる地図UIの拡大もしくは縮小の操作数
必要操作数
　入手したい情報にたどり着くまでに必要となる, コンテンツ検索以後の手操作数
　(従って最初のキーワード検索を除く)

ったことが大きな理由である. この後, 語を追加した絞り込み検索を行い, 上位のコンテンツから内容を確認した

一方ケース iii の場合, 最初のコンテンツ検索の結果の先頭に出現したコンテンツが, 望んだ情報を持っていたため, 被験者は最短の操作数でたどり着いた. この結果は偶発的ではなく, 話題の特徴に依存すると考えられる. "グローバルコミュニケーション"は, 元来国境を越えて情報の共有を図ろうとするものであり, そのための社会的また情報技術的な課題とその解決策を議論している. 従って, 代表的なコンテンツに重要な地域や場所に関し述べられている可能性は高い.

これに対し"地理学"は, 特定の都市や地域を議論する学問ではないため, 重要な地域や都市を論じているコンテンツそのものが少ない. これにより, 従来手法を用いるとキーワードの追加による検索の絞り込みを行う必要があった.

一方, iv で我々が提案する機能を用いると 6 操作で必要な情報を獲得できた. これは最初のキーワード検索の結果に対象コンテンツが含まれており(例え下位にリストされていても), コンテキスト(今回は"学術関連地名"を持っているか)の篩にかかるため, 専用の GUI に配置されたためである.

5.3 専用 GUI の利用による緩やかな絞り込み

4 つのいずれのケースにおいても確認できたシステムの特徴は, 情報獲得に至る「緩やかな絞り込み」をコンテキストに従った視点で全体を「俯瞰」しながら実現できることである.

多くの場合, 検索されたコンテンツが, 利用者が望む情報(もしくはその部品情報)であるかどうかを判断するには, 検索結果として表示されたコンテンツの内容を一つずつ人手によって判別する必要がある. そしてこの手間をできるだけ削減するために, 検索用のキーワードを追加し, 出力件数の削減をおこなう. ところが, この操作は本来必要としている情報までも検索結果からはじき出してしまう過度な最適化を行うこともある.

我々の開発したコンテンツ編纂機能は, 一旦キーワード検索されたコンテンツに対し, 指定したコンテキストの用語(今回は学術関連地名)で篩にかけ, 残ったものだけを対象に専用の GUI (今回は世界地図)上に表示する. 利用者は, コンテキストに従ったコンテンツの並びを俯瞰しながら, さらに注目する部分に絞り込んでいくことが可能となる. このように我々の開発したシステムは, 水口の述べる必要な情報を獲得する際の過度な候補削減を避け緩やかな絞り込みを実現したと言える[水口 95].

6. まとめ

我々は, インターネットのコンテンツ検索に対する新たな需要が発生しつつあると考える. 具体的には, 検索したコンテンツを特定のコンテキストに従って並び替える機能であり, 検索されたコンテンツを俯瞰し, 偏りやコンテンツ同士のつながりと関係性を知ることで, 新たな発見や発想・着想の支援に貢献できると期待される. 我々は, この機能を"コンテンツの検索・編纂"と位置付けている.

今回の実験では, 4 つの話題に対し, 重要な意味を持つ地名や都市名及び地域名を知りたいというコンテキストを想定した. この実験を通し, 我々の提案の有効性を確認した.

7. 謝辞

本稿は, 文部科学省科学技術振興調整費の支援による研究の一部である.

参考文献

[山名 05] 山名, 村田:検索エンジンの概要, 情報処理 9, Vol.46 No.9, 2005

[水口 95] 水口, 増井, ボーデン, 柏木. なめらかなユーザインタフェースによる地図情報検索システム. インタラクティブシステムとソフトウェア III: 日本ソフトウェア科学会 WISS'95, 近代科学社, 1995

[Hoeber 06] O. Hoeber , X. D. Yang:A comparative user study of web search interfaces: HotMap, Concept Highlighter, and Google, In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2006.

[Yao 02] Y.Yao: Information retrieval support systems, In Procs. Of the 2002 IEEE World Congress on Computational I Intelligence, 2002