2C2-06

Weblogにおける文書作成支援のためのエゴセントリック検索

An Egocentric Search Method on Weblog for Authoring Support

沼 晃介[*1*2] 大向 一輝[*1*2] 濱崎 雅弘[*1*2] 武田 英明[*2*1]
NUMA Kosuke OHMUKAI Ikki HAMASAKI Masahiro TAKEDA Hideaki

[*1] 総合研究大学院大学
The Graduate University for Advanced Studies

[*2] 国立情報学研究所
National Institute of Informatics

In this research, we propose an egocentric search method and an authoring support system for Weblog. In the search method, the importance of the target information is measured by the distance between myself (ego) and the information. The proposed system finds the related contents to the editing document with the search method. We performed an experiment and found this distance is related to the similarity between documents.

1. はじめに

近年の情報技術の進展により，文書をコンピュータ上で作成することが一般的になってきた．これに伴い，情報技術を用いて文書作成を支援する研究や製品開発が多数行われている．代表的な例が，アウトラインプロセッサ，あるいは構造化エディタである．これは，文書の構造を可視化し全体の見通しを立てやすくすることなどによって，概念の形成や文章の構成，文の記述および修正を支援するシステムである．これら既存の文書作成支援システムは，主として個人の知識を整理することを目的としている．しかしながら，文書の作成においては，まず論旨に関連する情報を収集した後に，それらに新たな関係や新たな情報を加え，文書としての体裁を整えることが一般的である．最も創造的な活動は，新しい関係や情報を付加する部分であるが，その活動を行うには十分な関連情報の収集が必要である．

一方，Web はいまや，我々の生活に不可欠な情報源のひとつとなりつつある．しかし，既存の Web 情報検索手法には，大別して 2 つの問題がある．第一は，クローリングにかかるコストの問題である．多量の情報を 1 箇所に収集することは難しく，仮に収集することができたとしても，それらを常に最新の状態に保っておくことは困難である．第二にあげられるのは，レイティングの問題である．既存手法の多くは，文書の表層的な情報をもとに重要度評価を行うため，利用者の多様な要求に応えられるとは限らず，検索結果にノイズが入ることがある．

これらの問題に対処するため，本研究では，文書を作成する個人を取り巻く人と情報のネットワークに着目する．情報を創造し発信するのは人であるが，人は無から情報を生み出すことはできない．人は，周囲からの情報をもとに個人の中に概念を形成し，生み出した情報を再び社会に還元するという営みの繰り返しによって文明を築いてきた．個人の創造的活動を支援するにあたり，その個人を取り巻く人と情報を利用することは自然であると考えられる．

本研究では，Web における文書作成を対象として，作成中の文書に関連する他の文書を，作成者の周囲の人およびコンテンツのネットワークを利用して，検索および提示する手法を提案する． またこの手法のシステムへの実装および実験による提案手法の評価を行う．

連絡先：沼 晃介， 総合研究大学院大学， 〒101-8430 東京都千代田区一ツ橋 2-1-2， Tel: 03-4212-2681， Fax: 03-3556-1916， kosuke@numa.jp

2. エゴセントリック検索

本研究では，ユーザを取り巻く人間関係およびコンテンツ間の関係を利用した「エゴセントリック（自分中心）」な情報検索を提案する．エゴセントリック検索とは，「自分を中心としたネットワークにおいて，自分からの距離の近さに基づき情報の重要度を評価する情報検索」と定義される[Ohmukai 2003]．これは，ユーザの近くにあるコンテンツは，そのユーザにとって興味深い情報を含んでいるとの仮説に基づく．

エゴセントリック検索を用いた場合，例えば，図 1 のようなエゴセントリックネットワークがあったとき，自分との接続関係以外の条件がすべて同じであれば，ノード A はノード B より近くに存在するため，高く評価される．

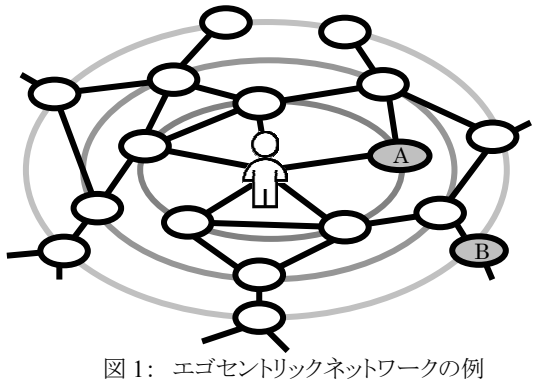

図 1： エゴセントリックネットワークの例

本研究では，エゴセントリックな情報検索を実現するため，Web における個人として，Weblog を利用する．

3. Web 上の個人としての Weblog

近年の Web において，Weblog（blog とも呼ばれる）という形式の Web サイトが注目されている．Weblog についての定義には諸説があるが，概ね個人が日記やメモなどといった小さな文書を蓄積していく形態の Web サイトの総称であると理解されている．本研究では，Weblog サイトに含まれるひとつひとつの文書をエントリと呼ぶ．Weblog 上のエントリは，他の Web サイトや Weblog サイト内の他の文書へのリンクを多く含むという特徴があるため，Weblog 間の関係を個人間の関係と考えると，人と人とが文書を介して接続されていると捉えることができる．

Weblog が今日流行している要因のひとつとして，Weblog ツールと呼ばれるコンテンツマネジメントシステム（CMS）があげられる．Weblog ツールは，ユーザが作成した文書を，あらかじめ設定されたテンプレートに従って HTML 形式に加工し公開する．これにより，従来の HTML によるマークアップと FTP によるファイルのアップロードと比較して，情報公開にかかるコストが劇的に低減されている．

また，多くの Weblog ツールでは，HTML による情報の公開と同時に，RSS フォーマット[RSS10 2001]によるメタデータの配信を行うことができる．これにより，エントリ本文に加え，文書の作成日時や筆者，文書のカテゴリなどの付随する情報を，機械可読な形式で取得することが可能である．

Weblog を特徴付けるもうひとつの機能として，TrackBack[Trott 2002]があげられる．TrackBack とは，Weblog のエントリ間における言及関係を明示する仕組である．言及した側と言及された側の Weblog ツールどうしが連携することにより，言及された側のエントリから，言及した側のエントリへのリンクを生成する．厳密な意味での逆リンクとは異なるが，おおよそ逆リンクの自動生成機能と捉えることもできる．この TrackBack により，逆リンク相当の情報が得られるため，コンテンツの周囲に絞った小規模なクローリングでも情報の言及および被言及関係が取得できる．

このように，Weblog を基盤とすることにより，機械可読なフォーマットで個人の知識の総体とみなせる情報を取得し，さらに TrackBack により各情報間の関連を取得することが可能となる．

4. Weblog における文書作成支援システム

Weblog における文書の記述の際に，エゴセントリックな手法を用いて関連する文書を検索し，提示する文書作成支援システムを実装した．このシステムは，ユーザの Weblog からリンクおよび TrackBack をたどり，その個人を取り巻くエゴセントリックネットワークを作成する．ユーザが新たに文書を作成する際や既存の文書を修正する際に，作成したネットワークを用いて関連文書を検索し，提示する．

図 2 に，システムの構成図を示す．提案システムは，ユーザの周囲の文書を収集するクローラ，収集した文書を蓄えるキャッシュデータベース，および収集したエゴセントリックネットワーク内の文書から関連文書を検索，提示する機能を持つ Weblog エディタから構成される．

関連文書の検索には，リンク検索およびキーワード検索を用いる．

リンク検索は，ユーザが作成している文書に含まれるリンクを利用した文書検索である．編集中の文書からのリンクおよび逆リンクをもとに以下の 3 種類のコンテンツを検索する．

a) 直接関係コンテンツ
　ユーザが作成している文書が参照する文書からさらに参照される文書コンテンツ，または作成中の文書を参照している文書をさらに参照している文書コンテンツを，直接関係コンテンツと呼ぶ．直接関係コンテンツによる検索手法を，Relative Chain Search という．

b) 共参照コンテンツ
　ユーザが作成している文書が参照する文書を参照している文書コンテンツを，共参照コンテンツと呼ぶ．共参照コンテンツによる検索手法を，Relative Co-reference Search という．

c) 共引用コンテンツ
　ユーザが作成している文書を参照している文書から同時に参照されている文書コンテンツを，共引用コンテンツと呼ぶ．共引用コンテンツによる検索手法を，Relative Co-citation Search という．

図 3(a)における，文書 A の直接関係コンテンツは，文書 C および E である．図 3(b)における，文書 A の共参照関係コンテンツは，文書 C である．また，図 3(c)における，文書 A の共引用関係コンテンツは，文書 C である．

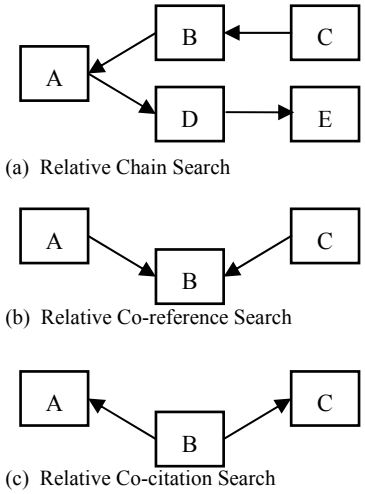

図 3： リンク検索

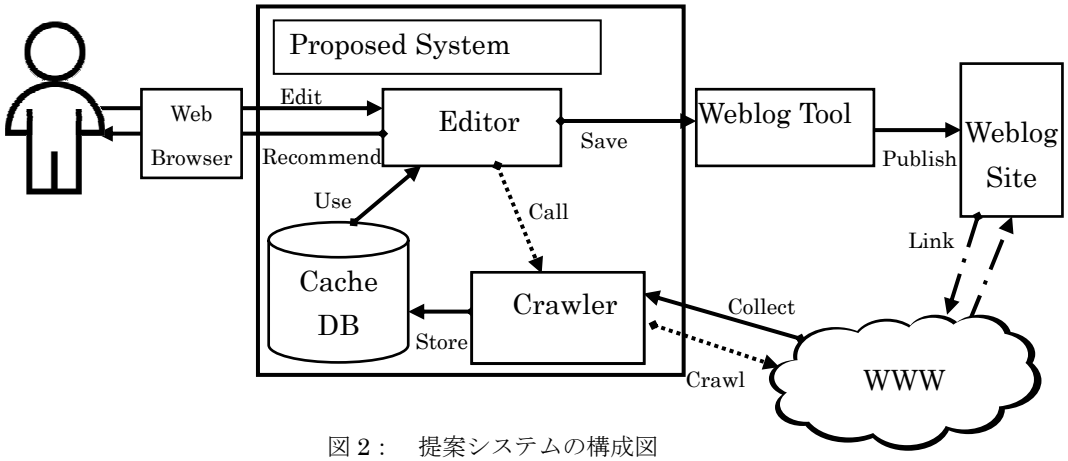

図 2： 提案システムの構成図

キーワード検索は，文書間の接続関係ではなく，文書コンテンツそのものを利用する検索である．キーワード検索では，ユーザ自身が指定した検索語を含む文書を，エゴセントリックネットワーク内から検索する．これは通常の Web 検索と同様に，能動的かつ明示的に要求する文書を検索する機能をユーザに提供する．発見された文書は，ユーザと文書の間の距離をもとに順位付けて提示する．

図 4 に，システムの動作画面を示す．ユーザは提示された文書を読み，参考になるものがあればその文書へのリンクを，作成している文書に追加することができる．この作業によって作成中の文書が充実するとともに，エゴセントリックネットワークも更新され，システムの検索結果が変化する．これらのプロセスの繰り返しによって文書の質のさらなるブラッシュアップが期待できる．

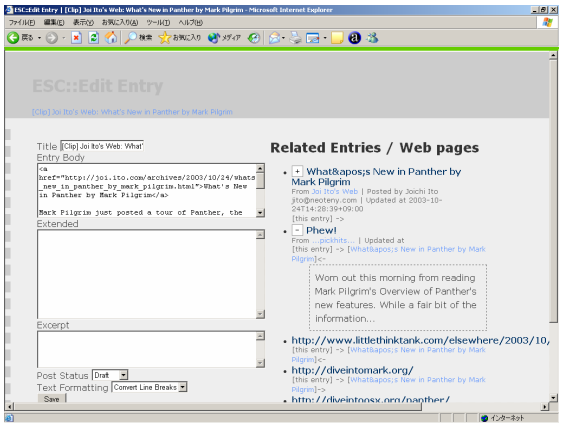

図 4： システム動作画面

5. 実験

ユーザからの距離が近い文書ほど有用であるという仮説を検証するため，実験を行った．文書の有用性を客観的な指標により計測することは難しいため，今回の実験では，ユーザの文書に類似した文書は有用性が高いという作業仮説を設定した．

5.1 実験の内容および手順

実験の内容および手順は以下の通りである．

1. クローラに，クローリングの始点となる Weblog サイトの URL を与え，この Weblog サイトを中心とするエゴセントリックネットワークを構築する．構築するネットワークは，距離 4 までの文書とした．
2. 収集したネットワーク内の文書を，距離に基づき分類する．ユーザを中心として，同一距離にある文書を，グループとした．今回の実験では，エゴセントリックネットワーク内のすべての文書が，距離 1 から距離 4 までの 4 グループのいずれかに分類される．
3. 同じグループどうしを含む，すべてのグループのペアについて，それぞれのグループ内に含まれるすべての文書の組み合わせで，類似度を計算する．

文書 a と文書 b の類似度は，以下の手順により計算される．

1. 文書 a および文書 b を，形態素解析し，出現する単語を抽出する．ここでは，名詞，複合名詞，記号，アルファベットを用いた．そのうち，代名詞など，文書の特徴抽出に貢献しない一部の語を，ストップワードとして除いた．
2. 抽出した単語列をもとに，文書のキーワードベクトルを作成する．
3. キーワードベクトルをもとに，TFIDF 法を拡張した SMART システム[Salton 1983]による類似度算出法をもとに，以下のように定義した類似度の値を計算する．SMART による文書 a をキーとした文書 b の類似度を *SMART(a,b)*とおいたとき，a と b の類似度 *similarity(a,b)* は，下式で与えられるものとする．

$$similarity(a,b) = \frac{\frac{SMART(a,b)}{SMART(a,a)} + \frac{SMART(b,a)}{SMART(b,b)}}{2}$$

SMART によるアルゴリズムでは，文書 a をキーとして文書 b の類似度を求めた場合と，文書 b をキーとして文書 a の類似度を求めた場合とで，類似度の値は異なる．また，とりうる値の範囲も文書によって異なる．そこで求めた値を，キーとなる文書自身と比較したときの類似度で割って正規化した値の平均をとる．

この実験において，形態素解析には茶筌[茶筌 1999]を，SMART のアルゴリズムによる類似度の計算には GETA[GETA 2002]を用いた．

実験に用いた Weblog サイトは，以下の 4 サイトである．

- サイト A
  http://www-kasm.nii.ac.jp/~numa/mt/
  活動の記録を，外部の Web ページおよび Weblog エントリへのリンクとともに記述する日記形式のサイトである．関連する Weblog エントリにもリンクを張っている．
- サイト B
  http://www-kasm.nii.ac.jp/~i2k/mt/
  活動の記録を，リンクをあまり使わずに記述する日記形式のサイトである．固有名詞などの具体的な情報は伏せられていることが多い．
- サイト C
  http://www-kasm.nii.ac.jp/~hamasaki/jimbo/
  活動の記録を，メモ書き的に記述する形式のサイトである．研究会等学術関係イベントの Web ページへのリンクを多く含む．各エントリの文章は短い傾向にある．
- サイト D
  http://www.semblog.org/i2k/
  研究プロジェクトに関する情報発信および意見提示を中心としたサイトである．特定のトピックに関する話題に特化した情報が記述されている．各エントリの文章は長く，多数のアウトリンクを含み，多数のトラックバックを受けている．

5.2 実験の結果および考察

各 Weblog サイトから取得したエゴセントリックネットワークに含まれる文書数を，距離別に表 1 に示す．エゴセントリックネットワークに含まれる文書数の差は，始点となる Weblog サイトが他の Weblog のエントリにどの程度リンクしているかが強く影響している．また，サイト B を始点として取得したエゴセントリックネットワークのネットワーク図を図 5 に示す．図の中心付近の円で囲ったノードが始点である．

表 1: 取得したネットワーク内の距離ごとの文書数

	サイト A	サイト B	サイト C	サイト D
距離 1	27	15	15	16
距離 2	22	4	12	62
距離 3	978	30	38	893
距離 4	1938	56	39	2187
合計	2965	105	104	3158

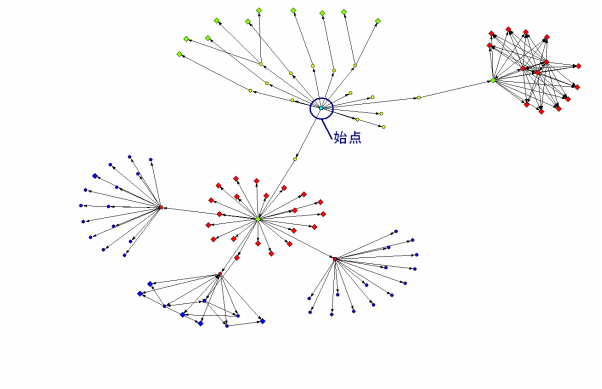

図 5: 取得したエゴセントリックネットワークの例 (サイト C)

それぞれのネットワークについて, 距離 1 の文書と, 距離 2, 3, 4 それぞれに含まれる全文書とを, 全組み合わせ総当たりにより類似度を計算し, 対象となる距離ごとにそれら類似度の平均を求めた. 距離と類似度の平均の値を表 2 に示す. また, そのグラフを図 6 に示す.

表 2: 自分の文書集合と距離 n の文書との類似度の距離別の平均値

	距離 1	距離 2	距離 3	距離 4
サイト A	0.037718	0.022829	0.013550	0.008218
サイト B	0.045831	0.018311	0.018788	0.021548
サイト C	0.028928	0.016824	0.009034	0.007248
サイト D	0.071033	0.052013	0.020000	0.017885

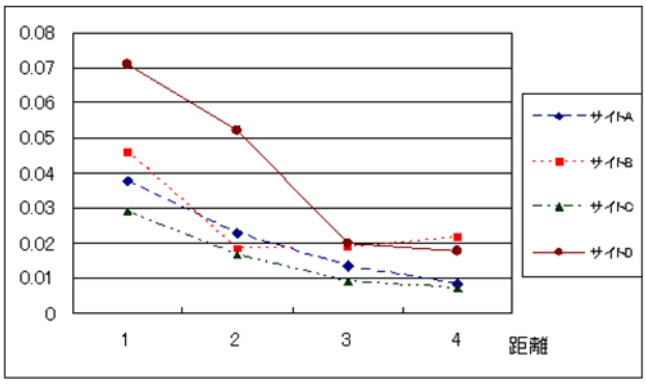

図 6: 自分の文書集合と距離 n の文書との類似度の距離別の平均値

類似度の値は概ね距離が大きくなるにつれ減少する傾向にあった. 今回の実験は, サンプルとなるサイトが 4 サイトと少ないため, 統計的に結論を導くことはできないが, 概ねエゴセントリックネットワークを構築した際, ユーザの興味に近いと考えられる文書がユーザの周囲に集まっているといえる.

各 Weblog サイトの他の文書へのリンクや被リンクの数や, リンク対象のサイトの情報量などによって, 収集されるネットワーク内の情報の量に大きな差があることが表 1 からわかる. 自分自身が作成したリンクのみならず, 自分からいくつかのリンクをたどった位置にある大きなサイトによって, 推薦されうる情報に変化が起こる. これは, たまたま自分からいくつかのリンクをたどった位置に多数の情報を蓄積した大きなサイトがあった際に, その先に接続される情報が内容的に発散してしまう可能性を示している. 集められた文書を作者ごとに整理することによって, 文書からのリンク数や情報量の影響を低減し, より質の高い推薦の手法を検討する必要がある.

また, 4 サイトの中では, サイト D が他のサイトに比べ, 文書間類似度が高くなっている. サイト D が特定のトピックの情報を扱っているため, 接続される文書においても関連するトピックが中心となっていると考えられる. この結果は, 文書のカテゴリ情報などを用いてトピックごとに整理しなおすことによって, より精度の高い推薦が行えることを示唆している.

6. まとめ

本研究では, 文書作成の支援を目的とした関連文書検索のために, エゴセントリックな情報検索手法を提案した. エゴセントリック検索とは, 自分を中心とするネットワークを築き, この上での「自分」と対象情報との距離を重要度評価の尺度に用いる検索手法である.

実験の結果, 中心に近い情報ほど, ユーザ自身の記述した文書に類似している傾向が確認された. これは情報とユーザの距離を情報検索に利用することの有効性を示している. しかしながら, 今回用いたエゴセントリックネットワークの作成手法では, まだ情報の整理が不十分である. より有益な情報を提示するためには, さらなる実験および分析を重ね, 情報検索および提示に最適なエゴセントリックネットワークの作成手法および, その上での距離の計算手法を検討する必要がある.

参考文献

[Ohmukai 2003] I.Ohmukai, K.Numa, H.Takeda: Egocentric Search Method for Authoring Support in Semantic Weblog, Workshop on Knowledge Markup and Semantic Annotation (Semannot2003), Held in conjunction with the Second International Conference on Knowledge Capture (K-CAP2003), 2003.

[RSS10 2001] RDF Site Summary 1.0 Specification Working Group: RDF Site Summary (RSS) 1.0, http://web.resource.org/rss/1.0/spec, 2001.

[Trott 2002] Benjamin Trott, Mena Trott: TrackBack Technical Specification, http://www.movabletype.org/docs/mttrackback.html, 2002.

[Salton 1983] Gerard Salton, Michael J. McGill: Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[茶筌 1999] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸: 日本語形態素解析システム『茶筌』version 2.0 使用説明書 第二版, NAIST Technical Report, NAIST-IS-TR99012, 1999.

[GETA 2002] 高野明彦, 丹羽芳樹, 西岡真吾, 岩山真, 今一修, 久光徹: 汎用連想計算エンジン "GETA", http://geta.ex.nii.ac.jp/, 2002.