人工知能学会研究会資料
SIG-SWO-A401-06

Weblog におけるエゴセントリック検索の提案と実装

Proposal and Implementation of an Egocentric Search Method on Weblog

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

本稿では，Weblog における文書作成のための情報検索および提示システムについて述べる．まず，関連文書の検索手法としてエゴセントリック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，ドキュメント距離およびサイト距離のいずれの手法でエゴセントリックネットワークを作成した場合も，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と考えられる．

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．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．

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[1]．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（egocentric）」という語は，社会ネットワーク分析における，エゴセントリックネットワークという語から借用したものである．社会学におけるネットワーク分析では，行為者の行為を決定するのは，行為者を取り囲む関係構造であると考える[2]．ネットワークを分析する際，分析の対象者を中心としたエゴセントリックネットワークに注目することにより，特定の行為者がその周囲にどのようなネットワークを作っているかを調べ，行為者間の関係を分析することができる．一方，行為者が結ぶネットワークの全体像をもとにしてネットワーク構造を分析する際，このネットワークをソシオセントリックネットワークという．そこで本研究では，Google[1] などに代表される従来の一般的な Web 検索のように，プロファイルや近傍情報を用いない，大規模かつ客観的な情報検索をソシオセントリック検索と呼び，提案手法と区別する．

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．

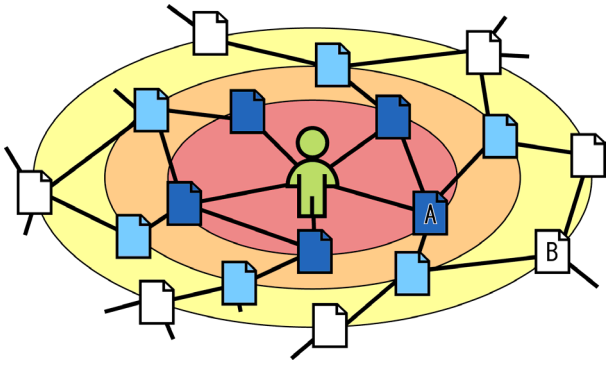


図 1： エゴセントリックネットワークの例

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．

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， numa@grad.nii.ac.jp

1) http://www.google.com/

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章を，あらかじめ設定されたテンプレートに従って HTML 形式の文書に加工し公開する．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[3]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，機械可読な形式で利用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[4]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．ここで，検索対象とする文書とは，Weblog のエントリならびに Weblog に含まれないひとつの HTML ページであるとする．

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．編集中の文書に直接接続された文書に加え，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

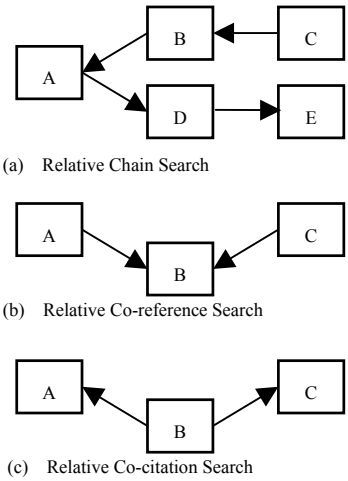


図 3：リンク検索

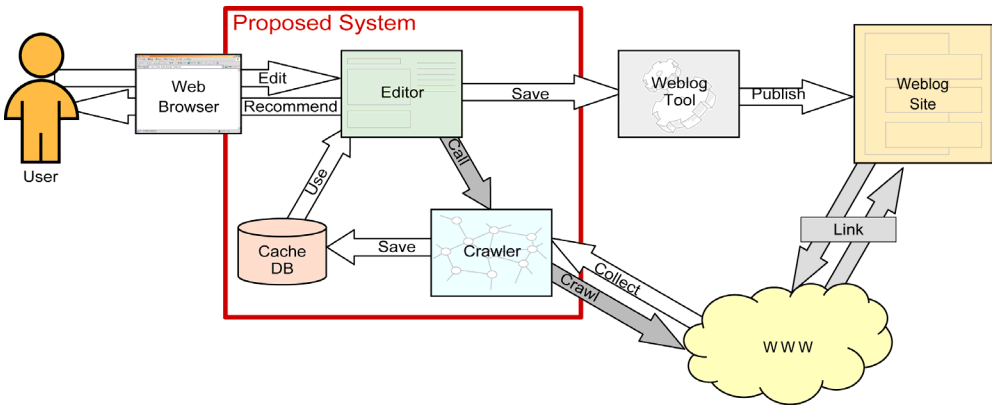


図 2： 提案システムの構成図

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)において作成中の文書 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．これらは，文書 A から直接接続されている文書 B などより弱いながらも，文書 A と関連があると考えられるため，あわせて提示する．

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．しかし，発見された文書を，ユーザと文書の間の距離をもとに順位付けて提示することにより，Google などのようなソシオセントリック検索とは異なる検索結果を提示することができる．

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．

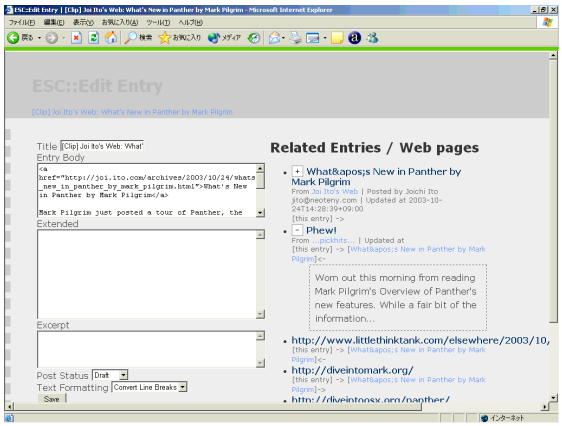


図 4： システム動作画面

5. エゴセントリックネットワークの作成

提案システムにおいて，精度の高い検索結果を提示できるかは，どのようなエゴセントリックネットワークを作成するかに大きく依存している．エゴセントリックネットワークの作成を行うクローラの動作について述べる．

5.1 クローラの動作

クローラは，始点となる Weblog の URI を与えたとき，順にリンクをたどって Weblog ページおよび Web ページを収集するプログラムである．以下のいずれかの条件が満たされたときに，探索を終了する．

1. 事前に設定された距離まで探索したとき
   ユーザはあらかじめ，自分からどの程度の距離までを探索するかを設定する．ここで用いる距離については後述する．
2. 非 Weblog ページが 2 つ続いたとき
   非 Weblog ページでは，情報が定型化されて提供されないため，事前にどのような情報であるかを予測することはできない．したがって，Weblog のエントリから直接関係コンテンツとなっている文書までを探索対象とし，それ以上の探索は行わない．

ここで Weblog ページと非 Weblog ページの判別には，対象ページに RSS によるメタデータが付加されているか否かの情報を用いる．すなわち，RSS のあるサイトは Weblog であるとみなし，RSS のないサイトの文書は一般の Web ページであるとする．RSS の発見は，RSS Auto Discovery[5]により行う．この手順による Weblog の分類結果は，必ずしも現実に一致しない．しかし本研究においては，その文書が実際に Weblog ツールによって作成されたかどうかという事実より，RSS ファイルを持つことによってサイトの範囲および文書の作成者が特定可能であるという状態を重視する．したがって，ここで起こりうる誤判断は問題としない．

ページ間の接続関係には，以下のものがある．クローラは，これら 4 つの関係を利用し距離を計算しながらリンクをたどる．

- Link
  Web ページ間がハイパーリンクにより接続されている順方向のリンク関係
- TrackBack
  TrackBack 機構を用いた逆リンクの関係
- Contain
  Weblog のトップページから，その Weblog が含む各エントリページへの関係
- Belong
  Weblog のエントリページから，そのエントリが属する Weblog のトップページへの関係

5.2 ドキュメント距離とサイト距離

エゴセントリックネットワークにおける距離には，ノードに相当する情報の粒度により，大別してドキュメント距離およびサイト距離の 2 通りの求め方が考えられる．

ドキュメント距離は，Weblog のエントリや，Weblog に属しない一般の Web ページなど，ひとつひとつの HTML 文書をノードとして構成されるネットワークにおけるノード間の距離である．Link，TrackBack，Contain，Belong のいずれの関係であっても，隣接するノードをたどる際に距離が 1 ずつ増加する．

サイト距離は，Weblog サイトをノードとするネットワークにおけるノード間の距離である．Weblog サイトは，個人を情報の編集主体とする Web サイトであるため，サイトをノードとするネットワークは，文書の言及関係に基づくパーソナルネットワークであるとみなすことができる．サイト距離は，サイトをまたがる Link および TrackBack をたどる場合にのみ，1 増加する．サイトをまたがらないリンクの場合や，Contain および Belong 関係の文書接続では，両文書が同一のサイトに属するため，同一の距離を与える．

ドキュメント距離は，文書情報そのもののつながりに基づく距離と捉えることができる一方，サイト距離は，情報の編集および発信の主体である個人間の距離を表している．

6. 実験

我々はこれまでに，エゴセントリック検索の有効性を，文書間の類似度により検証する実験を行った[6]．この実験は，ドキュメント距離に基づくエゴセントリックネットワークを用いた実験であ

った．今回我々は，前回の実験により収集されたものと同範囲の Web ページをサイト距離に基づき整理しなおし，エゴセントリックネットワークの作成手法間で，情報の整理にどのような違いが現れるかを調査した．この 2 つの手法の比較を明確にするため，前回の実験の概要から述べる．

6.1 ドキュメント距離に基づくエゴセントリック検索の有効性に関する実験

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，以下のような実験を行った．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，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．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 URI 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，ドキュメント距離 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．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文書のすべての組み合わせで，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[7]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)*といおいたとき，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[8]を，SMART のアルゴリズムによる類似度の計算には GETA[9]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．

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，ドキュメント距離別に表 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，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．赤い丸で囲まれた点が，始点となるノードである．

表 1: 文書によるエゴセントリックネットワーク内のドキュメント距離ごとの文書数

	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

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．距離ごとの類似度の平均値をグラフにしたものを図 6 に示す．

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および被リンクの数や，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，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．

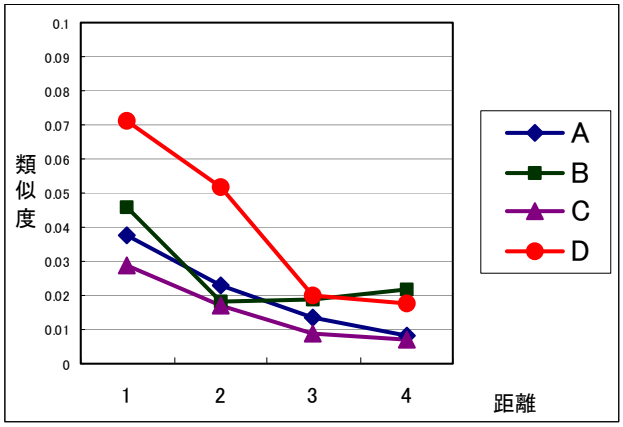


図 6: 自分の文書集合とドキュメント距離 n の文書との類似度の距離別の平均値

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．

6.2 サイトを単位としたエゴセントリックネットワークの縮約

先のドキュメント距離による簡易的な実験により，エゴセントリック検索が有効であることがわかった． だが，今後より精度の高い検索を行うには，より効果的なエゴセントリックネットワークの作成手法を検討する必要がある．今回我々は，ドキュメント距離に基づき収集した文書群を，サイトを単位として整理しなおすことにより，文書に基づくエゴセントリックネットワークとサイトに基づくエゴセントリックネットワークの比較を行った．

ドキュメント距離によるクローラを用いて収集した文書のキャッシュ内において，始点となる Weblog からサイトごとに文書をグループ化し，ネットワークを再構成する．ここで単位となるサイトとは，以下の手順により導き出される．

HTML 文書が，Weblog のトップページあるいはエントリページである場合，その文書は Weblog のトップページに代表されるサイトに属しているとする．Weblog ページであるかどうかは，RSS の有無により判別し，Weblog のトップページの URI はその RSS の channel 要素内の link 要素から取得する．

HTML 文書が Weblog に属していない場合，その文書がおかれている Web サーバのドメインを単位として，グループ化を行い，そのグループをサイトとみなす．非 Weblog サイトにおいて，一般的にサイトの単位を抽出することは困難である．ひとつの Web サーバに複数の Web サイトが置かれていることも多い．加えて RSS などによりメタデータが記述されていない場合，文書の筆者を同定することは，より難しい．本研究では，サイトの背後にいる，文書の編集者たる個人に注目するため，こうした情報が取得できない文書を大雑把にひとまとめにすることに大きな問題はないと考えられる．

サイト間は，サイト内に含まれる文書間が接続関係にあるとき，接続されているとする．

これらの手順によりサイトごとにノードを縮約し構成されたエゴセントリックネットワークのノード数を，表 2 に示す．括弧内は各距離に含まれるサイトのうち，Weblog サイトの数である．また，構成され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 7 に示す．赤い丸で囲まれた点が，始点となるノードである．

表 2: サイトによるエゴセントリックネットワーク内のサイト距離ごとのサイト数とそのうちの Weblog サイト数

	サイト A	*サイト B*	*サイト C*	*サイト D*
	サイト(Weblog)	サイト(Weblog)	サイト(Weblog)	サイト(Weblog)
距離 1	1 (1)	1 (1)	1 (1)	1 (1)
距離 2	10 (5)	4 (2)	11 (0)	11 (4)
距離 3	313 (128)	11 (4)	26 (3)	212 (102)
距離 4	606 (162)	0 (0)	4 (0)	510 (124)
合計	930 (296)	16 (7)	42 (4)	734 (231)

ドキュメント距離での実験と同様の手法により，距離 1 の文書と，距離 2，3，4 それぞれに含まれる文書とを，全組み合わせ総当りにより類似度を計算し，対象となる距離ごとにその平均値を求めた．距離ごとの類似度の平均値をグラフにしたものを図 8 に示す．

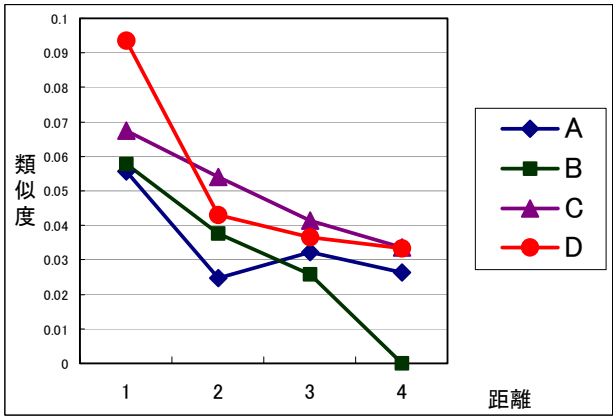


図 8: 自分の文書集合とサイト距離 n の文書との類似度の距離別の平均値

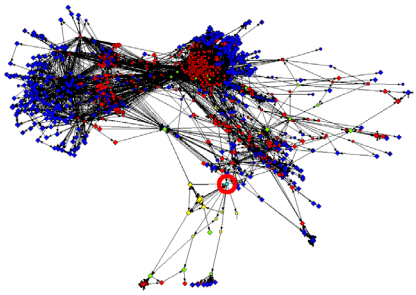
文書を単位とする場合の実験同様，サンプルとなるサイトが少ないため， 統計的に結論を導くことはできないが，類似度は距離に伴い減少する傾向が見られる．

ドキュメント単位およびサイト単位で作成したネットワークにおける，ノードの数およびリンクの数を表 3 に示す．

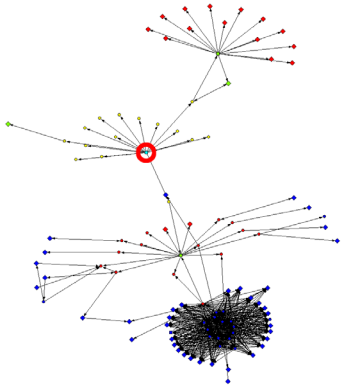
表 3: エゴセントリックネットワークに含まれるノード数およびリンク数

	サイト A		*サイト B*		*サイト C*		*サイト D*	
	文書	サイト	文書	サイト	文書	サイト	文書	サイト
ノード数	2966	930	106	15	105	16	3159	734
リンク数	23053	2260	1120	27	320	27	79174	2097

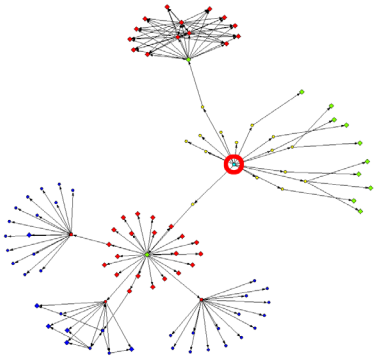
サイトを単位としてネットワークを縮約することにより，平均してノード数はおよそ 1/5，リンク数はおよそ 1/25 まで減らすことができた．



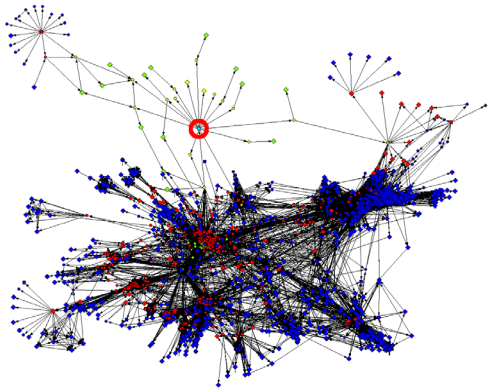
(a) サイト A



(b) サイト B

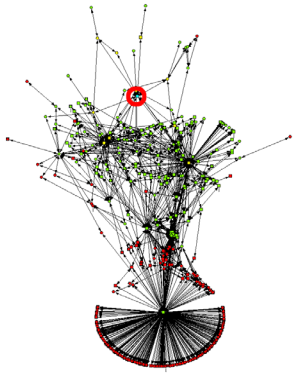


(c) サイト C

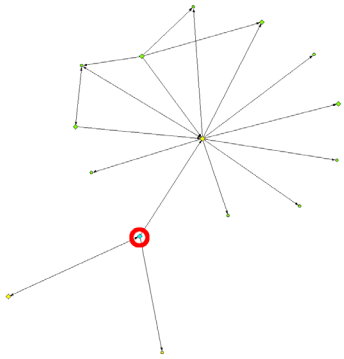


(d) サイト D

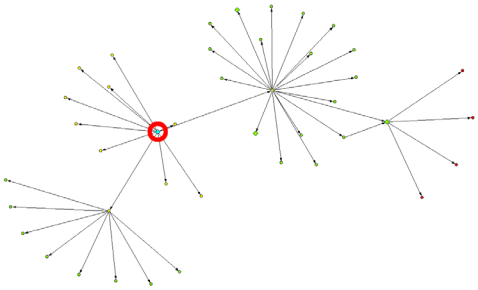
図 5： 文書をノードとするエゴセントリックネットワーク



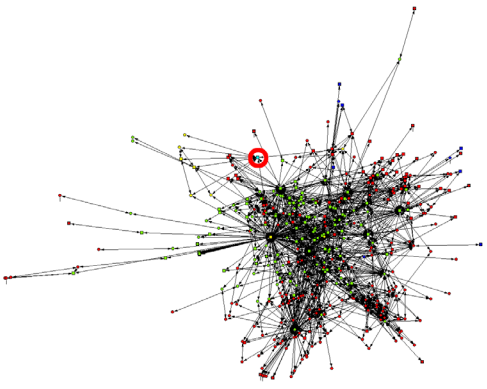
(a) サイト A



(b) サイト B



(c) サイト C



(d) サイト D

図 7： サイトをノードとするエゴセントリックネットワーク

サイト単位に分類された各文書の，サイト距離とドキュメント距離の対応を，サイト別に図 9 に示す．同じサイト距離に含まれる各文書をドキュメント距離ごとにカウントし，そのサイト距離に含まれる割合を示したものである．

同じ文書は，サイト距離でネットワークを構成した場合，ドキュメント距離の場合に比べて 1～2 程度近くなるものが多い．これはサイト内のリンクにおいて距離が増加しないためである．一方，ドキュメント距離よりサイト距離が遠くなる場合もある．これは，始点となる Weblog サイトのトップページに対し，他サイトから直接リンクを張った場合である．

実際には，サイト距離 3 および 4 には，ドキュメント距離が 5 以上のものが含まれる．しかしながら，今回の実験では，あらかじめドキュメント距離を 4 までに指定し取得した文書群を，サイトに基づき整理しなおしたため，ドキュメント距離が 5 以上となる文書は考慮されていない．サイト B において，サイト距離 4 となる文書が存在しないのは，ドキュメント距離 4 までに含まれるすべての文書が，サイト距離 3 以内に対応したためである．したがって図 9 が完全な対応を表しているとはいえない．同様の理由により，図 8 に示したサイト距離での文書類似度も，距離 3，4 の値は正確ではない．これらを厳密に調査するには，より大きなドキュメント距離の文書までを収集する必要がある．

以上の実験により，ネットワークのノード数，リンク数，距離の観点から，エゴセントリックネットワークをサイトに基づき縮約することにより，ネットワーク構造が簡潔になるといえる．このことは，図 5 および図 7 のネットワーク図を比較することからもわかる．

7. 今後の課題および発展性

実験の結果，エゴセントリックネットワークを利用することにより，ユーザの興味に近い情報が，ユーザの近くに現れる傾向がわかった．しかしながら，実験の対象としたサイト数が少ないため，一般性のある統計データを得ることができなかった．今後，より多くの Weblog サイトを用い，距離を伸ばしたエゴセントリックネットワークを構築することにより，詳細な分析を行う必要がある．

また，システムの効果を確かめるには，被験者を用いた実験を行う必要がある．今回の実験では，ユーザの Weblog と，エゴセントリックネットワーク内の文書の類似度に基づき興味の近さを推定した．文書の有用性は類似度に現れるとの仮定に基づいていたが，実際にユーザの創造活動にどの程度寄与するかは確かめられていない．同時に，エゴセントリックな情報検索の結果と，ソシオセントリックな検索の結果を，ユーザがどのように捉えるかを比較することにより，両検索の補完関係を検討する．

実験の結果，単純にハイパーリンクによる接続関係を利用した場合，リンクしたサイトによって構築されるエゴセントリックネットワークに大きな差が生まれることがわかった．リンク対象が多数のリンクを含んでいた場合，ユーザの興味に関連しない文書が，ユーザの近くに現れる可能性がある．このためには，文書のリンクの数や，文書のカテゴリ分類情報などにより，リンクに重み付けをし，適切な距離を求める手法を考える必要がある．

また，人と人，あるいは人と文書との接続関係について，検討を進める．今回は，文書間の言及関係をもとにして，接続関係を抽出した．今後は，他の手法によるユーザ間の接続関係と組み合わせることを考える．FOAF[10]は，人間関係を記述する

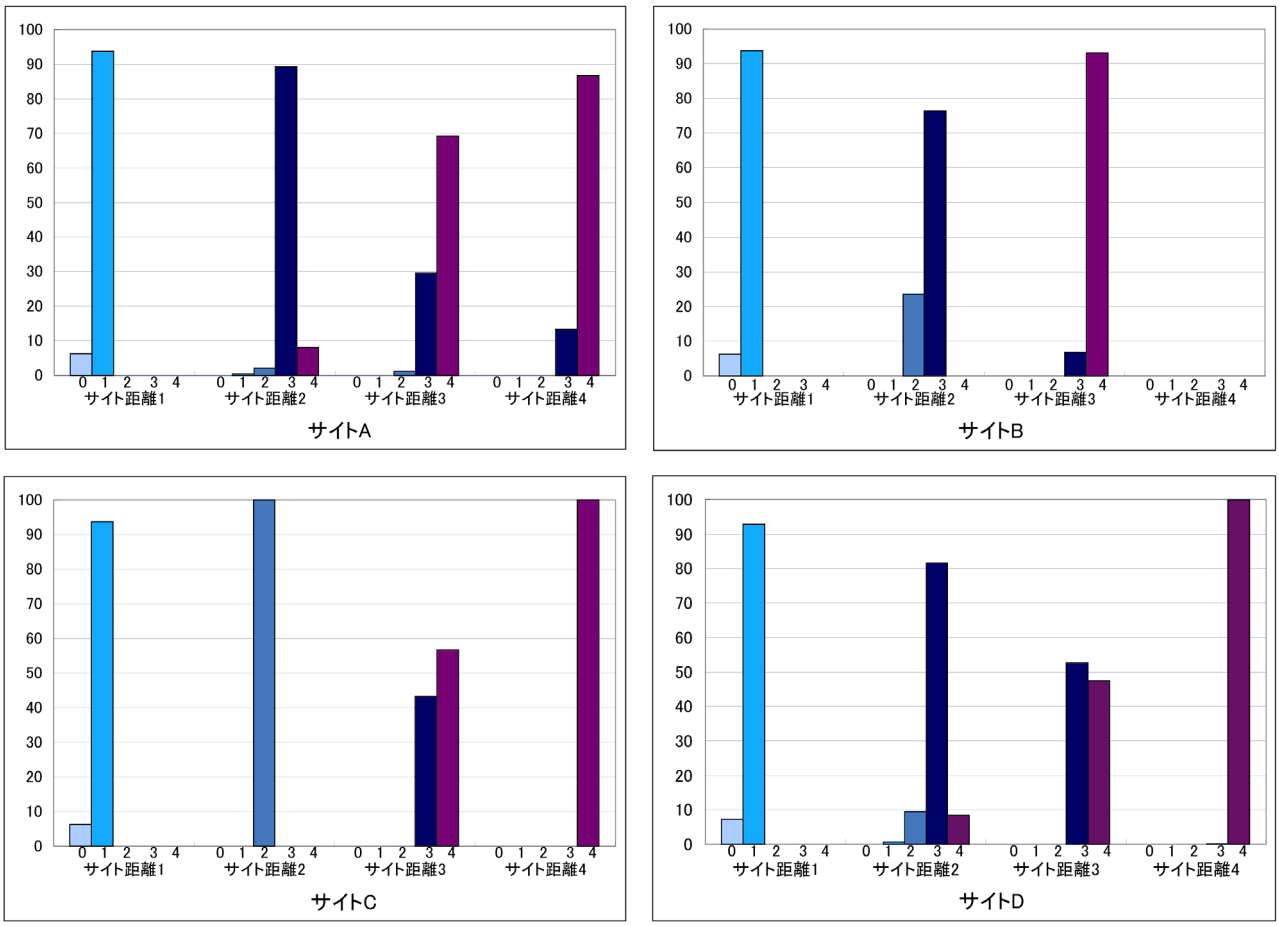


図 9： サイト距離とドキュメント距離の対応

ためのメタデータフォーマットであるが，こうした文書を介さない，ユーザ間の直接的な人間関係との併用手法を検討する．

以上の課題を検討した上で，ユーザにとって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，距離の定義を検討する．

ユーザが作成もしくは編集を行おうとする文書のコンテキストに沿った情報を，ユーザの興味を反映した評価に基づき提示することが可能となれば，ユーザの情報収集活動を通じてユーザの創造的活動を支援することができると考えられる．

8. 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，ドキュメント距離およびサイト距離のいずれの手法でエゴセントリックネットワークを作成した場合も，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．今後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していく．

参考文献

[1] 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.

[2] 安田雪:『社会ネットワーク分析―何が行為を決定するか―』, 新曜社, 1997.

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

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

[5] Mark Pilgrim: RSS auto-discovery [dive into mark], http://diveintomark.org/archives/2002/05/30/rss_autodiscovery, 2002.

[6] 沼晃介，大向一輝，濱崎雅弘，武田英明: 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，第 18 回人工知能学会全国大会論文集, 2C2-06, 2004.

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

[8] 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.

[9] 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.

[10] Dan Brickley, Libby Miller: the 'friend of a friend' vocabulary, http://xmlns.com/foaf/0.1/, 2003.