Wikipediaを用いたコミュニティ型コンテンツの概要抽出に関する研究

Extraction of the Outline of a Community type content by using Wikipedia

灘本　明代[*1]　荒牧　英治[*2]　阿辺川　武[*3]　村上　陽平[*4]
Akiyo Nadamoto　Eiji Aramaki　Takeshi Abekawa　Yohei Murakami

[*1]甲南大学 Konan University
[*2]東京大学 The University of Tokyo
[*3]国立情報学研究所 National Institute of Informatics
[*4]独立行政法人　情報通信研究機構 National Institute of Information and Communications Technology

It is difficult to grasp the outline of community type content such as Blog, SNS and message board, because multiple users write the part of the content freely. In this paper, we propose the system which compares a thread of community type content with a content of wikipedia. We focus on the table of content of a wilipedia. The system compares the comment of the thread with a content which is divided a table of content of a wikipedia, and represent the similar table of content of wikipedia. Then, user can grasp the outline of community type content when he/she brows the similar table of content of a wikipedia.

1. はじめに

現在 Blog や SNS 等コミュニティを中心としたコンテンツがインターネット上には多数存在する．我々はこのようなコミュニティにより作成されたコンテンツをコミュニティ型コンテンツと呼ぶ．コミュニティ型コンテンツは複数の人々が自由に記述することにより生成されたコンテンツであるため，一つのテーマで書かれた発言の集合であるスレッド（例えば mixi ではトピックを指す）の概要を一目で把握するのは困難である．また，コミュニティに依存してコンテンツが作成されているため，偏った発言である場合が多い．コミュニティの参加者にとって，自分達のコミュニティがあるテーマに対してどの部分について議論されているのか自分達の立ち位置を見たいとか，また投稿する際に過去の発言はどのような発言があったのか一目で見たいという欲求を満足する事ができないのが実情である．一方，Wikipedia は複数の人々の記述により生成されたコンテンツではあるが，中立的な観点が Wikipedia の基本的な方針の一つ[*1]であり，その点がコミュニティ依存のコミュニティ型コンテンツとは大きく異なる．そこで我々は，あるコミュニティの一つのスレッドと，そのスレッドが対象としているテーマの Wikipedia の記事とを比較することにより，そのスレッドの概要がわかるのではないかと考え，Wikipedia を用いたコミュニティ型コンテンツの概要抽出手法を提案する．具体的には，Wikipedia の記事の目次 (TOC) 構造に注目し，その目次構造を構成するセクションとコミュニティ型コンテンツのあるスレッドの各の発言とを比較し，類似しているセクションを示す目次タイトル（索引語）を提示することにより，そのコミュニティ型コンテンツの概要の一覧とする．この時，対象となるコミュニティ型コンテンツのスレッドのテーマと Wikipedia の記事が一対一の対応に成っているとは限らなく，つまりはテーマに対してコンテンツの粒度が異なる場合ある．そこで，対象となる Wikipedia の記事の周辺の記事も対象として比較することを考える．以下の手順にてコミュニティ型コンテンツの概要抽出を行う．ここで本論文では，コミュニティ型コンテンツにおいて，各ユーザが記述した最小単位のコンテンツを発言と呼び，一つのテーマで書かれた発言の集合をスレッドと呼び，あるテーマで統一されたスレッドの固まりをそのテーマのコミュニティと呼ぶ．また，Wikipedia の目次の最小単位が指し示すセクションのコンテンツを Wikipedia の最小コンテンツと呼ぶ．

1. ユーザが入力したキーワードから，Wikipedia の記事及びコミュニティ型コンテンツのスレッドを提示する．対象となるコミュニティ型コンテンツのスレッドが複数存在する場合，ユーザにより一つのスレッドが選択される．
2. 対象スレッドからキーワードを抽出する．
3. キーワードを用いてスレッドを構成する各発言と Wikipedia の最小コンテンツを比較する．
4. Wikipedia の記事を解析し，その周辺のページを取得する．
5. Wikipedia の周辺記事を対象として (3) の比較を行う．
6. (5) の類似度が閾値以上の場合，その周辺のページもスレッドの概要を示す対象コンテンツとする．
7. (3) と (6) の結果，目次を色づけをして提示する．

以下，2 章では関連研究を，3 章ではコミュニティ型コンテンツのスレッドと Wikipedia の記事との比較方法を，4 章ではプロトタイプシステムについて，そして 5 章では実験についてのべ，6 章でまとめについて述べる．

2. 関連研究

Wikipedia に関連する研究は多数あるが中山 [中山 08], Suchanek[Suchanek 07], Wu[Wu 08], Gabrilovich[Gabrilovich 07] らに代表されるように Wikipedia から知識を抽出し利用する研究が数多い．これらの研究は

連絡先: 灘本明代，甲南大学知能情報学部，
〒 658-8501 兵庫県神戸市東灘区岡本 8-9-1，
E-mail:nadamoto@konan-u.ac.jp

*1 http://ja.wikipedia.org/wiki/中立的な観点

Wikipedia のカテゴリ構造やリンク構造を用いて知識を抽出しているのに対し，本論文では Wikipedia の記事の目次構造を用いて同じテーマを持つコミュニティ型コンテンツのスレッド概要一覧を抽出し提示している点が異なる．また，川場ら [川場 08] はあるトピックに有用なブログサイトを検索する応用例として Wikipedia を用い，Wikipedi の記事に対応したトピックのブログサイトを検索している．堀ら [堀 08] はユーザのクエリからその意図に沿った拡張クエリを作成する際に Wikipedia を用いるシステムを提案している．これらの研究はクエリの拡張に対して Wikipedia を用いているが，本論文ではユーザの示したスレッドからキーワードを抽出し，そして Wikipedia を用いてスレッドの一覧を示す点が異なる．

3. Wikipedia とコミュニティ型コンテンツとの比較

3.1 キーワードの抽出

Wikipedia の最小単位のコンテンツとコミュニティ型コンテンツの発言は両方ともコンテンツの分量が少ない場合が多い．このようにコンテンツの分量が少ない場合，文書全体の単語を対象として比較しその類似度を求めると適合率が悪いことが以前の我々の研究で判明している [灘本 08]．

これまで我々は ComparativeWeb[灘本 03] において小山ら [小山 02] の提案する TopicStructure の改良版を使用して，複数の Web ページを比較してきた．本研究では，この TopicStructure をコミュニティ型コンテンツに対応し，それを比較を行うためのキーワードとする．スレッド P における Topic Structure$T_s(P)$ は主題語の集合 Sub_p と内容語の集合 Con_p の 2 つの組からなる．すなわち，$T_s(P)$ は以下のとおりである．

$$T_s(P) = (Sub_p, Con_p)$$
$$Sub_p = \{s_1, .., s_m\}$$
$$Con_p = \{c_1, .., c_n\}$$

主題語

主題語はコミュニティ型コンテンツのスレッドのタイトルを構成する単語である．名詞のみを対象とする．ここで，コミュニティ名に含まれる単語はそのスレッドの特徴を示している単語ではないと考えストップワードとする．

内容語

指定したスレッド内全体 SD_n を対象とし，ある単語 C_{mn} の重み CW_{mn} がある閾値 α 以上の単語を内容語とする．単語は名詞のみを対象とする．ここで CW_{mn} は TF・IDF 法を用いる．

3.2 コンテンツの比較

図 1 に示すように，TopicStructure を用いてコミュニティ型コンテンツの発言毎に Wikipedia の目次を構造する最小コンテンツと比較し類似コンテンツを抽出する．TopicStructure を構成する単語 i の文書 D_j における重み W_{ij} は以下の式で示すように TF・IDF 法を用いる．ここで D_j はコミュニティ型コンテンツの一つの発言または Wikipedia の最小コンテンツとする．

$$W_{ij} = tf_{ij} \times idf_j \quad (1)$$

$$idf_j = \log \frac{N}{df_j} \quad (2)$$

上記 TopicStructure の各単語の重みを求めた後，コサイン相関値を用いて類似度計算を行う．類似度がある閾値 β 以上の Wikipedia の最小コンテンツをその発言の内容が含まれて

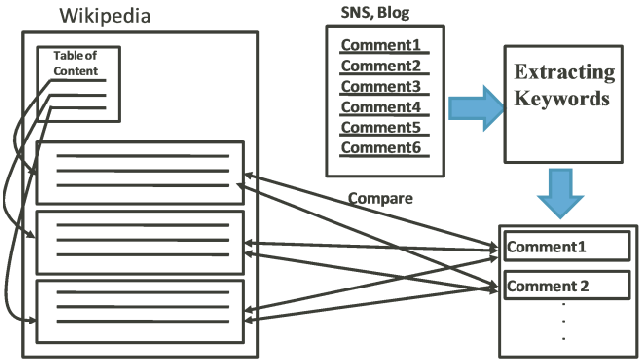

図 1: 比較イメージ図

いるコンテンツであるとみなし，その最小コンテンツを指す目次の索引語をスレッドの概要を示す単語とする．この時，一つの最小コンテンツがある発言に対し，TopicStructure の主題語，内容語両方とも類似している場合，主題語のみ類似している場合，内容語のみ類似している場合がある．主題語，内容語両方とも類似している場合は最もその発言の特徴を示していると考え，これら 3 つのパターンを色分けして表示することにより，概要を一目でより把握しやすいようにする．

3.3 Wikipedia の周辺記事の取得

コミュニティ型コンテンツのスレッドと Wikipedia の記事とを比較する時，対象となるコミュニティのテーマについて各のコンテンツの粒度が異なる場合がある．そこで，我々はコミュニティのテーマと同じ Wikipedia の記事の周辺の記事も比較対象とすることにより粒度の問題を解決することを考える．ここで周辺記事とは，ある記事の上位概念を示している記事と関連を示している記事及び下位概念を示している記事とする．本論文では，周辺記事を取得する手法の最初のアプローチとして上位概念及び関連記事の取得方法について述べる．

上位概念

単語 X の上位単語の発見手法について述べる．その手法は下記の仮定に基づいている．

1. 単語 X が,「X は〜である.」もしくは,「X とは〜である.」という文章に存在する場合,「である.」の直前の単語は X の上位単語 Z である．
2. Wikipedia に記載されている記事は世間一般的な単語を網羅している．またそれらの記事間には概念の上下関係が存在する．
3. リンクタグ（$< a >$ 〜 $< /a >$）に囲まれる文字列は 1 つの単語もしくは，複合語としての意味を持つ．
4. Wikipedia 内で，ある単語 X に関して説明文の始めの 1 文は（1）の形式をとる可能性が高い．

1 つめの仮定について，単語 X と Z とは is-a 関係にあることは中山ら [中山 08] の研究よって証明されている．この場合，”単語 X は Z である ”と言い換えることができる．2 つ目の仮定は，Wikipedia が従来の研究において単語間の概念構造を構築する手法に数多く利用されている．また数々の研究によりそれらの有用性が示めされている．3 つ目の仮定では，リンクタグのそのアドレス先には必ず情報が存在することが容易にわかる．4 つ目の仮定においても，同じように仮定（1）の文脈が存在することが最も多い．我々は上記の仮説に基づいて，

図 2: プロトタイプシステム初期画面図

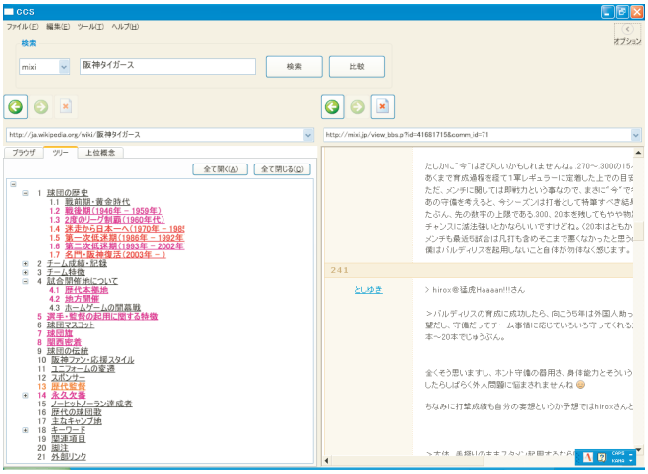

図 3: プロトタイプシステム解析結果図

上位概念の記事の抽出に以下の提案を行う．まず，単語 X を Wikipedia にクエリとして渡す．大多数の Wikipedia の記事は，単語 X の記事の冒頭部の 1 文は「X は〜.」という文が存在する．我々はこの Wikipeda の構造に注目し，この冒頭の 1 文を取り出す．そして，この文の文末に最も近いリンクタグで囲まれた文字列は，単語 X の上位単語である場合が多いため，そのリンク先のページが上位概念を示す記事であるとする．

関連記事

Wikipedia の記事には関連項目がある場合が多い．そこで，この関連項目に掲載されているアンカーのリンク先を関連記事とする．

4. プロトタイプシステム

提案手法を用いてプロトタイプシステムを C#にて開発した．プロトタイプシステムのフローを以下に示す．図 2 にプロトタイプシステムの初期画面を，また図 3 に解析結果を示す．

1. ユーザは比較したいコミュニティのテーマをキーワードとして入力する．
2. ユーザの入力したキーワードからそのキーワードのコミュニティのサイトのリストと Wikipedia のページを検索し，コミュニティのリストを右画面に，Wikipedia を左画面に表示する（図 2 参照）．
3. ユーザは (2) で表示されたコミュニティのリストから概要一覧を見たいコミュニティの一つのスレッドを選択する．
4. システムはユーザの指定したスレッドからキーワードを抽出する．
5. システムは抽出したキーワードを用いてユーザが指定したコミュニティ型コンテンツのスレッドの各発言と Wikipedia の目次毎の最小コンテンツとを比較する．
6. Wikipedia の目次の階層構造を利用して，キーワードの主題語と内容語両方が類似している目次を赤で，主題語のみ類似している目次をオレンジで，内容語のみ類似している目次をピンクで，主題語内容語ともに類似していない目次を黒字で表示する．(図 3 参照)

5. 実験

種々のコミュニティのテーマを用いて提案手法の有用性を計り，提案手法の問題点を抽出する実験を行った．ここではキーワードの有用性及び目次構造を利用する事の有用性を検討するために，比較対象の Wikipedia の記事は，スレッドのテーマの記事 1 ページのみとし，上位概念，同位概念を示すページの比較は対象外とした．さらに，キーワードに TopicStructure を用いて類似度計算をした場合と，キーワードを特に用いず，対象のコンテンツの名詞すべてを対象として類似度計算をした場合との比較実験も行った．実験に用いたコミュニティのテーマは以下の方針で選んだ．

- 固有名詞のうち，組織名と個人名との比較
  固有名詞をコミュニティ型コンテンツのテーマとしている場合，その固有名詞は有名な組織や個人である場合がほとんどであり，Wikipedia に掲載されている可能性が高い．そこで，話題が広義な会社や団体等の組織名とそれと比較して話題が狭義な個人名とを比較する．組織名と個人名とで個人名の方が狭義の話題になり，Wikipedia を用いることが有効であると予測する．
- 上位クラス，下位クラスの比較
  テーマの構造が上位クラスの場合とその下位クラスの場合とを比較する．つまりは提案手法は Wikipedia を用いているため，よりインスタンスに近いテーマの方が有効であると予測するが，その比較実験を行う．例えば，JAL がテーマの場合，JAL をテーマとしているコミュニティとその下位クラスであると考えられる JAL のサービスの一部であるマイレージ（JAL マイレージバンク）をテーマとしているコミュニティにおいて，どちらが Wikipedia を用いることにより有用なのかを比較する．

表 1 に一部の結果の適合率を示す．ここで求めた適合率は以下の通りである．

$$\text{適合率} = \frac{\text{類似度が閾値以上の目次項目の内正解の項目数}}{\text{類似度が閾値以上の目次項目数}} \times 100 \quad (3)$$

表 1: 評価実験に用いたテーマとその結果

対象テーマの説明	対象とするテーマ名（クエリ）	TopicStructure 使用適合率 (%)	TopicStructure 不使用適合率 (%)
組織名と個人名			
組織名	JAL	42	36
組織名	阪神タイガース	43	28
個人名	相武紗季	53	42
個人名	金本知憲	61	61
上位クラスと下位クラス			
上位クラス	JAL	55	36
下位クラス	JAL マイレージバンク	83	81
上位クラス	ドコモ	53	38
下位クラス	ドコモダケ	63	61

5.1 考察

キーワードに TopicStructure を使用した場合と使用しなかった場合とは，すべてのデータにおいて，TopicStrucutre を使用した場合の方が適合率がよくなり，TopicStructure の有用性を示す事ができた．また，各データに関しての考察は以下の通りである．

- 固有名詞のうち，組織名と個人名との比較
  組織名と個人名とで個人名の方が狭義の話題になり，Wikipedia を用いることが有効であると予測したが，予測通りの結果となった．組織名のように広義の話題に対してより適合率を上げるための手法の検討が今後の課題となった．

- キーワードの上位クラス，下位クラスの比較
  ここでは，上位クラスは話題が広義であるため，下位クラスの方がより有効であると予測したが，予測した通りの結果となった．結果より，上位クラスをテーマとするコミュニティに対してもその下位クラスの項目の目次も有効であるということがわかり，下位クラスをどの範囲と決定するかが今後の課題となった．

6. まとめ

本論文では Wikipedia の目次を利用し，コミュニティ型コンテンツの一つのスレッドの概要の一覧を提示する手法を提案した．具体的には，コミュニティ型コンテンツのあるスレッドから主題語と内容語から成る構造化されたキーワードである TopicStructure を抽出し，その TopicStructure を用いてコミュニティ型コンテンツの発言と Wikipedia の目次構造からなる最小コンテンツとを比較し，類似した目次項目を抽出し，その類似した目次項目をわかりやすく提示するシステムの提案を行った．さらに Wikipedia の記事の周辺情報として上位概念，関連記事のページを抽出し，そのページの比較対象として，類似した目次項目を抽出し提示することを行った．

謝辞

本研究の一部は，平成 21 年度科研費特定領研究域「コミュニティ型コンテンツのコンテンツホール検索に関する研究」(課題番号：21013044，代表：灘本明代) 及び甲南大学平生太郎基金科学研究奨励助成金による．ここに記して謝意を表します．

参考文献

[中山 08] 中山浩太郎，原隆浩，西尾章治郎: 自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法, 電子情報通信学会データ工学ワークショップ (DEWS'08) 論文集, 2008 年 3 月.

[Suchanek 07] F.M.Suchanek, G.Kasneci, G.Weikum: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, Proceedings of the 16th International World Wide Web Conference(WWW2007), pp.697-706, Banff, Canada, May 2007.

[Wu 08] Fei Wu, Daniel S. Weld: Automatically Refining the Wikipedia Infobox Ontology, Proceedings of the 17th International World Wide Web Conference(WWW2008), pp.365-644, Beijing, China, April 2008.

[Gabrilovich 07] Evgeniy Gabrilovich, Shaul Markovitch: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, Proceedings of the International Joint Conference on Artificial Intelligence 2007(IJCAI 2007), pp.1606-1611, Hyderabad, India, Jabuary 2007.

[川場 08] 川場真理子，中崎寛之，宇津呂武仁，福原知宏: Wikipedia エントリとブログサイトの対応付けのための特定トピックのブログサイト検索, 電子情報通信学会データ工学ワークショップ (DEWS'08) 論文集, 2008 年 3 月.

[堀 08] 堀憲太郎，大石哲也，長谷川隆三，藤田博，峯恒憲，越村三幸: Wikipedia への関連単語抽出アルゴリズムの適用とその評価, 情報処理学会研究報告, Vol.2008, no.56, 2008-DBS-145, pp.81-88, 2008 年 6 月.

[灘本 08] 灘本明代, 荒牧英治, 阿辺川武, 村上陽平: Wikipedia を用いたコンテンツホール検索の提案，情報処理学会研究報告, Vol.2008, No.88 2008-DBS-146，pp.259-264 2008

[灘本 03] Akiyo Nadamoto and Katsumi Tanaka, " A Comparative Web Browser (CWB) for Browsing and Comparing Web Pages ", Proceedings of the 12th International World Wide Web Conference (WWW2003) , pp.727-735, Budapest, Hungary, May 2003.

[小山 02] 小山 聡, 田中 克己: 質問の階層的構造化を用いた Web 検索手法の提案, 日本データベース学会 Letters, Vol.1, No.1, pp.63-66, 2002 年 10 月