概念チェーンを用いた観光ブログからの意外性発見

曾, 駿
九州大学

中藤, 哲也
九州大学

殷, 成久
九州大学

呉, 小斌
九州大学

他

出版情報：情報処理学会研究報告. 2011, pp.1-4, 2011-03. 情報処理学会
バージョン：

情報処理学会研究報告
IPSJ SIG Technical Report

概念チェーンを用いた観光ブログからの意外性発見

曾駿[†]　中藤哲也[†]　殷成久[†]　呉小斌[†]　廣川佐千男[†]

観光地の評判やグルメ情報、イベント紹介など、多くの観光情報がブログで発信されている。しかし、ユーザにとって潜在的に価値ある情報が大量の情報に埋まれてしまうことがある。本発表では、概念チェーン(R. K. Srihari)を用いて、ユーザのキーワードと潜在的な関連がある意外な情報を発見する方法を提案する。

Concept Chain based Serendipitous Search from Tourism Blog

Jun Zeng[†] Tetsuya NakaToh[†] Chengjiu Yin[†]
Xiaobin Wu[†] and Sachio Hirokawa[†]

Blog has become a popular tool to share people's ideas including information of tourist attraction, gourmet restaurants and campaign. However, the large volumes of blogs have created the potential of a vast amount of valuable information buried in those corpus. This paper proposes a novel approach to discover potential relevant information based on Concept Chain, in order to provide a serendipitous search from tourism blog.

1. はじめに

近年、ブログの急速な発展に伴い、ブログで自分の感想、意見を公開する人が増加しつつある。通常の Web ページとは異なり、速報性、リアルタイム性のある新鮮な情報が発信されることから、Blog は注目される情報源になっている。米国 Technorati 社の調査によると [1]、2007 年 3 月時点で同社が追跡するブログ数は、世界全体で 7000 万以上、日本語による記事は全体の 37%であり、日本語ブログは世界一位の発信量と報告されている。更に、旅行業の発展により、観光地の評判やグルメ情報、イベント紹介など、多くの観光情報がブログで発信されている。しかし、ユーザにとって潜在的に価値ある情報が大量の情報に埋まれてしまうことがある。本論文ではこのような情報を意外性情報と呼び、そのための検索を意外性検索と呼ぶ。

一般的な検索エンジン（google, yahoo など）ではキーワードと関連の強いページを検索するため、これらの意外性情報は無視されてしまうこともある。しかし、これらの意外性情報は意味がないわけではない。特に観光情報を検索する時には、思いかけない情報の方が価値は高い。本論文はブログの観光情報に対して、意外性検索の方法を提案する。

本論文の構成について簡単に述べる。２節では意外性検索と概念チェーンに関する関連研究、３節では概念チェーンの手法を用いた意外性検索について述べる。４節では観光情報の収集について述べ、５節では観光情報の意外性検索を行い、検索結果を評価する。最後に６節で全体のまとめを行い、今後の課題を述べる。

2. 関連研究

意外性（serendipity）検索に関する研究はいくつかある。Andre ら [2]はどのような検索結果に意外性を感じるか分析している。この研究では、検索結果とキーワードの関連度及び検索結果の面白さ２つの観点で、評価を行っている。アンケートを通じて、面白くて関連度が低い検索結果に潜在的な意外性があることを示している。しかし、意外性に影響する要因についての分析はしていない。

R. Beale[3] はユーザに意外性のあるウェブページを推薦するシステムを提案した。このシステムはユーザのアクセス履歴を分析し、ウェブページの意外度を計算する。意

[†] 九州大学
Kyushu University

外度が最も高いページをユーザに薦める。

本論文で使う概念チェーンに関する研究の始まりとしては Swanson[4]の ABC モデルがある。ABC モデルとは「A は B に影響を与え、B は C に影響を与えるともには、A は C に影響を与える可能性がある」という仮設に基づくモデルである。ABC モデルは概念チェーンのプロトタイプである。

本研究と最も近い研究として、Wei Jin ら [5]による概念チェーンを用いたデータマイニング手法がある。この研究では、単語の共起頻度によって概念チェーンを生成し、二つの単語の潜在的な繋がりを発見する。Wei Jin らの研究では、与えられた二つ単語間の概念チェーンを発見する。一方、本研究では与えられたキーワードに繋がっている概念チェーンを生成し、その概念チェーンを通して意外性がある単語を発見する。また、検索対象を文書ではなく文とすることで、チェーンのステップごとにその両端の２つの単語を含む文が得られ、２つの単語の関連を解釈できる。

3. 概念チェーン

本来、概念チェーンの論文 [5),6),7)]では、文書集合を対象としていたが、本論文では文の集合を対象とする。

3.1 概念チェーンの定義

本研究の目的としては概念チェーンを通じ、概念間の潜在的な関連を発見し、意外性のある情報をユーザに推薦することである。本節は概念チェーンに関する定義を述べる

定義１：文の集合が与えられた時、その文の集合に対する概念チェーングラフとは以下のようにして構成される重み付きグラフ G(N, E)である。

- N：文の集合に現れる単語の集合。
- E：同一文書に共起する２つの単語 A、B の組からなるエッジの集合。
- 重み(weight)：二つの概念の関連度である。

概念 A と概念 B の重み $W_{A,B}$ は式（１）で計算する。

$$W_{A,B} = \frac{N_{A,B}}{N_A + N_B - N_{A,B}} \qquad (1)$$

$N_A(N_B)$は概念 A（概念 B）が文の集合に出現する頻度である。$N_{A,B}$ は概念 A と概念 B が共起頻度である。

例えば、表 1 のような文集合がある。

表 1　文集合の例

文１
熊本市の安政町に行ってきました。
文２
安政町商興会の新年会は紅蘭亭で開催されました。
文３
・・・

「熊本」をキーワードとすると、表 1 の文集合から生成した概念チェーングラフは図 1 のようになる。

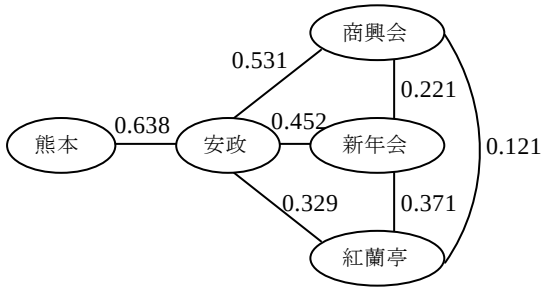

図 1　概念チェーングラフの例

図 1 の各エッジの重みは仮想値である。

3.2 概念チェーンのランキング

概念チェーングラフを生成した後に、各概念チェーンの優先順序を計算する。優先順序の定義は以下の通りである。

定義２：概念チェーン C_k の総辺数を $Length(C_k)$と表し、C_k の重みの和を $Weight(C_k)$ と表す。与えられた二つの概念チェーン C_i と C_j が以下の条件をいずれか満たす時、C_i は C_j より優先度上位にあるという。

- $Length(C_i) < Length(C_j)$；
- $Length(C_i) = Length(C_j)$ AND $Weight(C_i) > Weight(C_j)$

優先順序を算出したら、優先順序で各概念チェーンをランキングする。図 1 の概念グラフを例として、「熊本」と「紅蘭亭」間の各概念チェーンをランキングすると、表 2 のようになる。

表 2　「熊本」と「紅蘭亭」の概念チェーンのランキング

ランキング	概念チェーン
1	熊本 － 安政 － 紅蘭亭
2	熊本 － 安政 － 新年会 － 紅蘭亭
3	熊本 － 安政 － 商興会 － 紅蘭亭
4	熊本 － 安政 － 新年会 － 商興会 － 紅蘭亭
5	熊本 － 安政 － 商興会 － 新年会 － 紅蘭亭

定義３：$S=(C_1,C_2,\cdots C_n)$を与えられた２つの概念 A と B の概念チェーンとする。C_i（$C_i \in S$）が優先度で最上位になれば、C_iは A と B の最短概念チェーンという。

表 2 を例として、「熊本」と「紅蘭亭」の最短概念チェーンは「熊本－安政－紅蘭亭」である。

3.3 概念チェーンを用いた意外性検索

本節では概念チェーンで意外性検索の手順を述べる。

入力：キーワード K

処理１：K をルートとして、概念チェーングラフを生成する。

処理２：K と各概念の概念チェーンを優先度上位下位順でランキングする。

処理３：K と各概念の最短概念チェーンを抽出する。

出力：概念集合、最短概念チェーン集合、各概念ペアを含む文の集合。

4. 観光情報の収集

筆者らの先行研究 [8),9),10)] で利用した観光情報は、ブログではなく、九州観光推進機構で提供されている 312 件のイベント情報であった。本研究で取り扱うデータとして、1303 件のブログの記事を収集した。収集した情報は文単位で分割した。データの一部を表 3 に示す。

表 3　観光情報の例

```
. . .
340-55.txt
お昼は鹿児島の中心に移って、天文館地区のラーメン屋にて。
. . .
584-22.txt
刺身が少なかったですが、煮付（左端奥）やイカ刺し（右側奥）
もついて、品数は多いですね。
. . .
```

5. 検索事例

提案手法に基づく「意外性サーチエンジン」を作成した。サーチエンジンの画面を図 2 に示す。

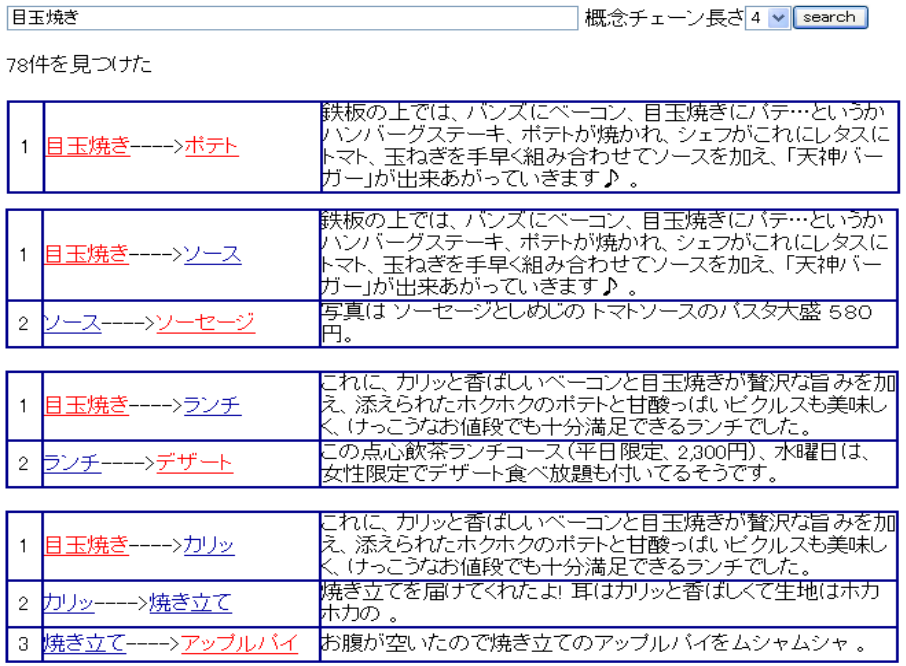

図 2　「意外性サーチエンジン」の画面

図 2 のように、テーブルで検索結果を分別した。テーブルごとに１つの概念チェーンを表す。テーブルの一行ごとに概念チェーンの各ステップの両端の２つの概念及びこの２つの概念を含む文がある。

表 4，5 の各行の先頭の単語が入力した検索語である。表 4 では、チェーンに現れる単語のほとんどが食べ物であり、意外な繋がりだが、納得できるものである。

表 4　意外性の概念チェーン

目玉焼き ― バーガー― グルメ ― クッキング ― コンビニ ― 握り飯
ホタテ ― スープ ― 煮込ん ― 甘辛く ― サンド ― ピラフ
コンニャク ― つくね ― コース ― おまかせ ― スペアリブ

しかし、表 5 の例では、容易に繋がりは推測できないパスも含まれるので、全体として納得しづらい結果となっている。

表 5　意外性のない概念チェーン

天神 ― やって来る ― しゅう ― 煮付け ― いわし
熊本 ― さわっ ― ハンドル ― なつかしい ― アイスキャンデー
九州 ― 取締役 ― (株) ― 小麦粉 ― バター

6. 結論と今後の課題

本論文では概念チェーンを用いた観光情報の意外性検索手法を提案し、「意外性サーチエンジン」を開発した。1303 件観光情報ブログに対する意外性検索を行った。しかし、検索結果に意外性がない結果も混在しているような欠点が明らかになった。

今後、我々はこの欠点の解決方法を考え、意外性検索の手法を改善するために、研究を行う予定である。

参考文献

1)　田中 憲光: 資料コーナー：ブログの実態に関する調査研究, 電学論D, Vol. 130, No. 5, pp.NL5_05 (2010).
2)　Andre P, Teevan J, Dumais ST: From X-Rays to Silly Putty via Uranus: Serendipity and its Role in Web Search, 27th Annual CHI Conference on Human Factors in Computing Systems, Boston, MA, USA ,APR 04-09(2009).
3)　R. Beale: Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing, Int. J. Human-Computer Studies 65, pp. 421–433 (2007).
4)　Swason, D. R. : Complementary Structures in Disjoint Science Literatures, Proc. of 14-th ACM SIGIR Conf. on Research and Development in Information Retrieval, ACM Press , pp.280-289 (1991).
5)　Wei Jin, Rohini K. Srihari, Xin Wu: Mining Concept Associations for Knowledge Discovery Through Concept Chain Queries, Lecture Notes in Computer Science, Vol.4426/2007, pp.555-562 (2007).
6)　Srihari, R., Ruiz, M., and Srikanth, M.: Concept Chain Graphs: A Hybrid IR Framework for Biomedical Text Mining. In Proceedings of the SIGIR Workshop on Text Analysis and Search for Bioinformatics (2003).
7)　Rohini K. Srihari, Sudarshan Lamkhede, Anmol Bhasin: Unapparent Information Revelation:A Concept Chain Graph Approach". CIKM '05 Proceedings of the 14th ACM international conference on Information and knowledge management, New York, USA, pp.329-330 (2005).
8)　X. Wu, S. Hirokawa, C. Yin, T. Nakatoh, Y. Tabata, Extraction and Comparison of Tourism Information on the Web, Proc. AROB2011(2011).
9)　殷成久, 呉小斌, 廣川佐千男, 中藤哲也: 観光イベントについての「といえば検索」の提案, 電子情報通信学会技術研究報告 110(301), pp.43-47 (2010).
10)　C. Yin, T. Nakatoh, S. Hirokawa, X. Wu, J. Zeng: A proposal of search engine “XYZ” for tourism events, Proc. JCAI2010 (2010).