Webマイニングにおける新たなアプローチ: 想定外検索(Search Engine Offering Result of No Assumption)の提案

New Approach in Web Mining: Proposal of Search Engine Offering Result of No Assumption

岡本 悠作[1,*]，澤野 弘明[1]，鈴木 裕利[2]，石井 成郎[3]，伊藤 誠[4]，原 崇[5]
Yusaku Okamoto，Hiroaki Sawano，Yuri Suzuki，Norio Ishii，Makoto Ito，Takashi Hara

[1] 愛知工業大学，[2] 中部大学，[3] 愛知きわみ看護短期大学，
[4] (株)名鉄情報システム，[5] (株)スナップショット
[1]Aichi Institute of Technology，[2]Chubu University，[3]Aichi Kiwami College of Nursing，
[4](c)Meitetsu Information System，[5](c)Snapshot
* x11030xx@aitech.ac.jp

Abstract

We propose a search engine offering a result of no assumption by web mining. A user searches a keyword, and a result including the keyword is obtained. However, the web page must have the keyword, and the user must know one. Our search engine provides a related web page by re-searching the obtained one. The proposed system uses Wikipedia since it is not hard to analyze the format. The 72% of the examinees answered positive feedback in the evaluation experiment, and the availability is indicated.

Keywords — Search engine, No assumption, Evaluation experiment, web mining

1. はじめに

ネットショッピングやSNS (Social Networking System) などのように，現実空間でのコミュニケーションが仮想空間に導入され，Internet初期に登場した静的なWebページの閲覧だけでなく，コミュニケーションツールとしてWebページが活用されている．またスマートフォンやタブレット端末の普及により，手軽に情報をWeb上に投稿することが可能であり，Webに含まれる情報量は拡大を続けている．IT専門調査会社IDC[1]は2020年までにはWeb上の情報量が40ゼタバイト (10^{21} byte)に達すると予想している．この情報爆発の影響を受け，いくつかの企業ではビッグデータの解析に基づいたマーケティング手法の提案，データ収集，営業コストの削減などの活動を積極的に進めている[2].

Web上から情報収集するには，キーワードによる情報検索が利用されることが多い．Google[3]のようなWeb検索エンジンでは，利用者の入力する検索キーワードに関連性の高いWebページ (リソース) を検索結果として提示している．その結果においては，検索キーワードとリソースとの関連性のみが評価されるため，間接的に関連のあるキーワードの提示が不可能であるという課題がある．すなわちキーワードを利用して検索する場合，Webページ内に必ず該当キーワードが存在するという観点から，利用者にとって既知の情報しか得ることができず，構造化されていないテキストデータから情報抽出するテキストマイニング[4]の分野では，利用者にとって新しい知見を提供できないという点で課題となっている．その課題に対して，新聞記事の文書集合を取得して階層ごとにクラスタリングし，トピックを関連させて分析者に提供するシステム[5]が橋本らによって提案されている．新聞記事の語彙に基づいてクラスタリングすることで，関連した記事を提供できるという点で有用であるといえるが，語彙同士の関連性を直感的に把握できないという課題がある．一方，キーワード検索時に関連性の深いキーワードを同時に提示する連想検索エンジンreflexa[6]がPreferred Infrastructureによって提供されている．また検索キーワードについて語られているブログから一緒に語られているキーワードの出現頻度でキーワードを提示するkizasi[7]が株式会社きざしカンパニーによって提供されている．これらの検索手法では利用者が，検索キーワードに関連した新たなキーワードの取得は可能であるが，検索キーワードとの関連性については明記されていない．

そこで本研究では，検索キーワードに関連のある事象を取得する方法として，検索キーワードのリソースが一定の規則で記述されているWikipediaに着目する．Wikipediaを利用したマイニングの研究[8]はいくつか提案されているが，上記で挙げている，関連キーワードとその関連性を提示する研究に関しては，筆者らが調査する限り確認されていない．本研究では，利用者が検索キーワードを入力して検索し，検索されたリソースに対してシステムが検索することで，入力された検索キーワードと共通語句で関連するキーワードを持つリ

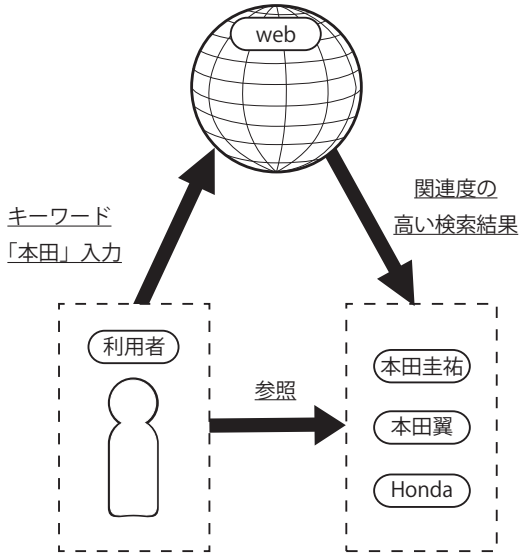

図 1 検索キーワードに基づく検索手法

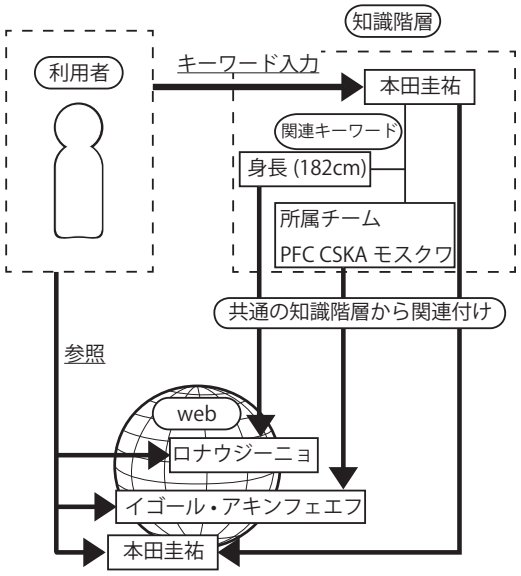

図 2 提案システムの検索手法

ソースを提示し，そのリソース間の関連性を同時に提供することを目指す．利用者にとって興味のある事象で検索した場合，提案システムで関連項目を取得するため，肯定的な想定外な検索結果が得られることが期待される．提案システムを定量評価するために，アンケート方式による実証実験を行い，提案システムの有効性を示す．

2. 提案システム

2.1 概要

GoogleなどのWeb検索エンジンでは，利用者が設定した検索キーワードに関連のあるリソースを提示している．例えば図1のように利用者が「本田」とWeb検索した場合，サッカー選手の「本田圭佑」やファッションモデルの「本田翼」や自動車メーカーの「Honda」のように,「本田」という語句が含まれるリソースが提示される．提案システムでは検索キーワードで得られるリソースから新たな検索キーワードとなりうる語句を抽出し，抽出語句を検索キーワードとしたリソースを利用者に提示する．抽出した検索キーワードと入力したキーワードが抽出語句に基づいて関連していることが特徴である．本研究ではこの共通事象である抽出語句を知識階層と呼ぶ．図2のように利用者が「本田圭佑」と検索した場合，知識階層として「サッカー選手の身長」が選択され，リソースとして「ロナウジーニョ」が表示される．提案システムによる検索では利用者の予想を超える結果が得られることを期待し,「想定外検索」と定義する．

この想定外検索を実現するために，検索キーワードの相関を分析する対象と，情報源の著作権問題の発生からの回避が必要である．この二つの問題の解決としてWikipediaの情報を利用する．WikipediaとはWeb上で百科事典を作成するプロジェクトであり，構成される記事の項目はすべて事象で構成される．またWikipediaの記事はWikiの形式で記述されることから，事象の説明のために書かれた名詞が，他の記事と結びつき，Webリンクされている．そのため，事象同士が関連づけ合う関係であることも有益である．そのため本研究ではWikipediaを情報のリソースとする．

2.2 構成する機能

提案システムは図3に示されるようにデータベース構築部，知識階層構築部，リソース表示部の機能から構成される．以下にそれぞれの特徴を示す．

データベース構築部 提案システムでは検索キーワードと知識階層に基づいて2回検索が必要なため，オンラインで取り扱うには処理時間が必要であるため，高速化のために予め提案システムで取り扱うデータベースを構築する．

知識階層構築部 Wikipediaでは記事の要約された規定フォーマットとして，Infoboxと呼ばれる基本情報が用意されている．図4にマルティン・ルターのInfoboxの表示例を示す．InfoboxはWiki形式で図5の記載方法でまとめられており，提案システムではInfoboxを知識階層として利用する．Infoboxから知識階層となる情報の抽出を行い，共通の情報を持つキーワードをグループ化して，データベースに格納す

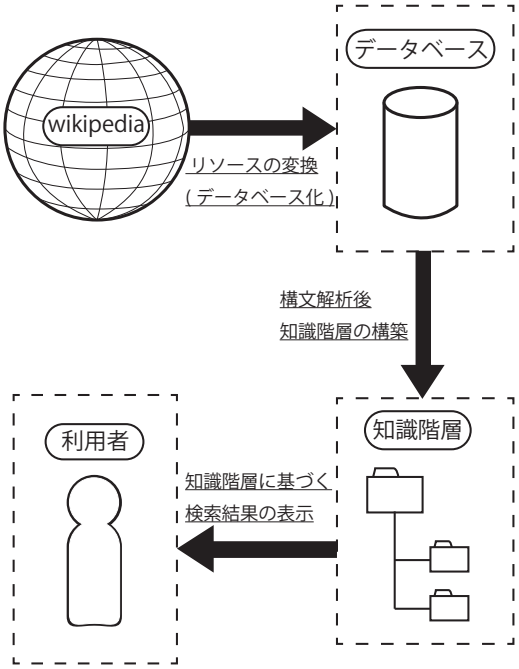

図 3 提案システムの構成

る.「本田圭佑」の例では，サッカー選手および身長が同じリソースがクラスタリングされる.

リソース表示部 知識階層構築部において知識階層に基づいてマイニングした結果を利用者に表示する．検索キーワードと出力されたリソースを関連付ける知識階層を提示するために，グラフ表示およびテキスト表示を行う．提案システムではテキストボックスに利用者が検索したいキーワードを入力後，検索ボタンの押下により検索結果が表示される．画面上のグラフ表示において利用者が入力したキーワードの知識階層は，灰色で表示され，共通の知識階層となる情報を持つキーワードを関連キーワードとして線分で結んで表示する．またテキスト表示における最大表示件数を利用者に入力させることで表示件数を制限させる.

3. 実験と考察

3.1 実験環境

提案システムで使用したライブラリを表1に示す．mongDB[1]は関係データベース管理システム (RDBMS: Relational Database Management System) 系ではないNoSQL（Not only SQL）に分類されるオープンソースのドキュメント指向データベースであり，RDBMSのように固定的なスキーマを定義しなくてもよい点，配列データ等も格納できる点，軽量なデータ記述言語であるJSON (JavaScript Object Notation) に似た形式でデータの格納が可能である．また，処理するスクリプト言語としてはRuby[2]を使用し，Webアプリケーションフレームワークとしては Rails（Ruby on Rails）[3]を導入する．Ruby におけるmongoDB用のドライバ（Ruby-mongo-driver）[4]を利用し，DBを取り扱う．そして，オブジェクトとリレーショナルデータベースの対応付けを行うO/RMapper としてMongoid[5]を使用する．提案システムに，データベースに変換されたリソースを知識階層に基づいてマイニングした結

[1] http://www.mongodb.org/

図 4 マルティン・ルターのInfobox表示

```
{{Infobox 人物
|氏名=マルティン・ルター
|画像=Martin Luther by Lucas Cranach der Ältere.jpeg
|画像サイズ=200px
|画像説明=[[ルーカス・クラナッハ]]画（[[1529年]]）
|生年月日=[[1483年]][[11月10日]]
|生誕地=[[神聖ローマ帝国]]、[[アイスレーベン]]
|没年月日=[[1546年]][[2月18日]]（62歳）
|死没地=[[神聖ローマ帝国]]、[[アイスレーベン]]
|職業=[[神学者]]、[[司祭]]、[[牧師]]
|配偶者=カテリーナ・ルター
|署名=Martin Luther Signature.svg
|署名サイズ=160px
}}
```

図 5 マルティン・ルターのInfobox記載方法

[2] https://www.ruby-lang.org/ja/
[3] http://rubyonrails.org/
[4] http://api.mongodb.org/ruby/current/
[5] http://mongoid.org

表 1 利用したライブラリ及びライセンス

ライブラリ	Version	目的	ライセンス
MongoDB	2.4.8	NoSQLデータベース	GNU AGPL v3.0
Ruby	1.9.3	Ruby処理形	2-clause BSDL Ruby's license
Ruby on Rails	4.0.2	Ruby WEBフレームワーク	MIT license
Ruby-mongo-driver	1.9.2	Ruby MongoDB Driver	Apache License, Version 2.0
Mongoid	4.0.0	Ruby MongoDB O/RMapper	MIT license
arbor.js	0.91	HTML5でグラフ表示	MIT license

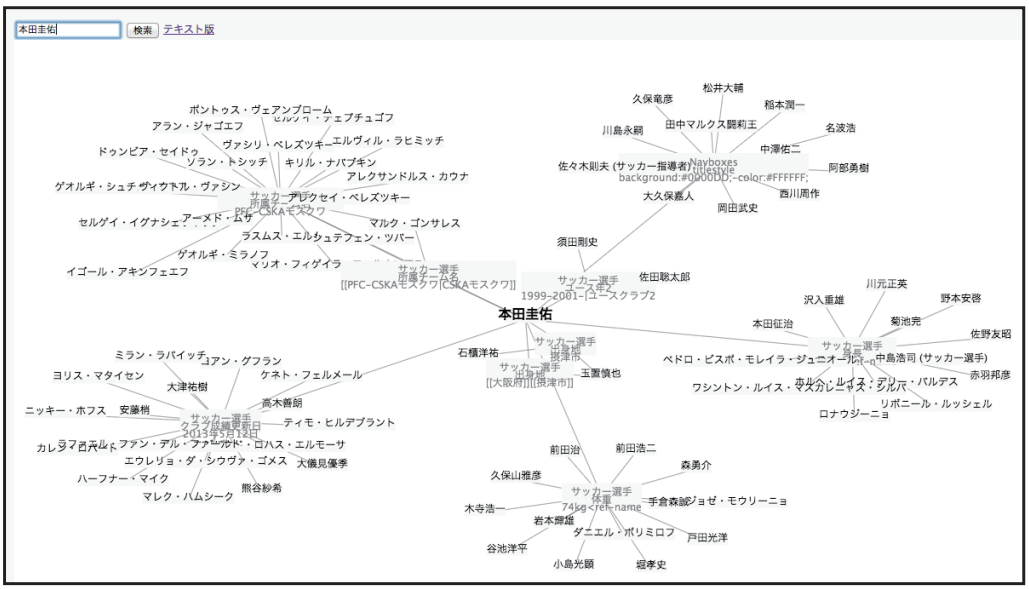

図 6 「本田圭佑」でマイニングをした結果(グラフ表示)

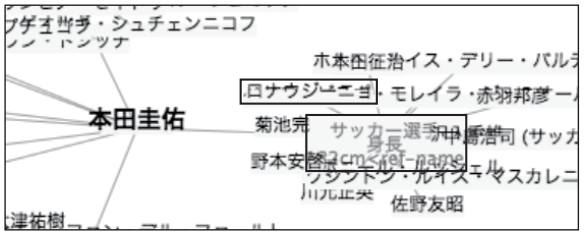

図 7 「本田圭佑」でマイニングをした結果
(グラフ表示)の一部拡大

果を表示可能とする出力機能を実装するために，HTML5 においてグラフ表示を可能とするライブラリである arbor.js[6]を使用する．なお，Wikipedia のデータベースは2013年12月20日に生成した．

3.2 実験

検索キーワードにサッカー選手である「本田圭佑」を入力として，提案システムでグラフ表示で検索した結果とその拡大図を図6, 7に示す．提案システムで取得した知識階層の一部を以下に示す．

- サッカー選手身長182cm
- サッカー選手体重74kg
- サッカー選手クラブ成績更新日2013年5月12日
- サッカー選手所属チーム名PFC-CSKAモスクワ

得られたリソースとして知識階層で身長が共通しているブラジルのサッカー選手「ロナウジーニョ」等が確認された．また，図8のテキスト表示では，グラフ表示では密集したリソースにより閲覧が困難なキーワードも，一覧表示により視認性が向上することが確認された．しかし，提示される情報量の多さから，利用者が求める情報の直感的な取得が困難である問題点も示された．一方で，「ロナウジーニョ」や所属しているチーム名等，執筆時[7]の情報とは異なる情報が提示されているという結果が得られた．現行システムではある時点のWikipediaのデータで構築されていることが直接的な原因であるが，Wikipediaの記事が不定期に不特定多数のユーザにより編集されているため，データベース構築時のデータ取得タイミングに依存するという問題点を抱えている．そのため更新頻度を短くすることによって情報の鮮度はある程度保つことが運用における解決方法であるといえる．

[6] http://arborjs.org/

[7] 2014年7月22日時点

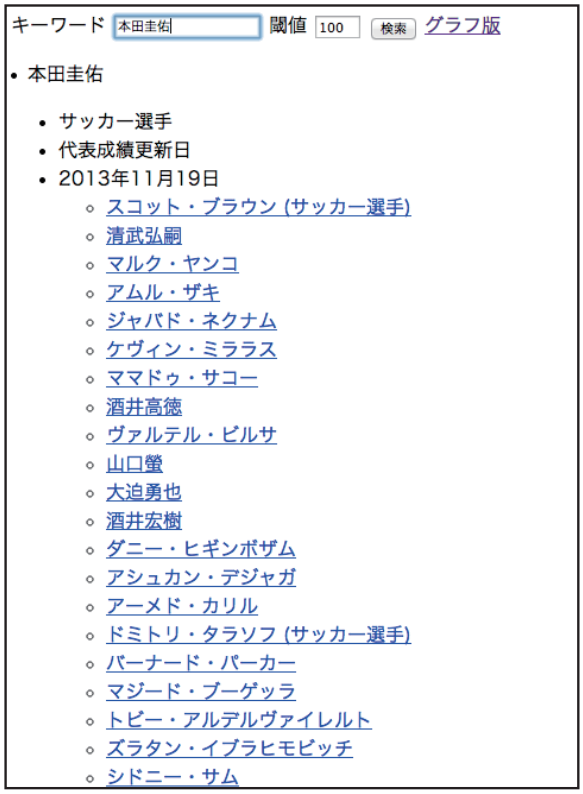

図 8 「本田圭佑」でマイニングをした結果 (テキスト表示)

また情報提示の問題として，戦国大名「徳川家康」のように知識階層の項目や出力リソースが膨大である場合，情報過多によって閲覧に耐えうる表示ができない結果が確認された．出力結果の知識階層や検索リソースの数によって，表示項目を限定する対応が今後の課題として挙げられる．

3.3 定量的評価

提案システムを定量評価するために，20歳から32歳までの男女26人の被験者に対してアンケート調査を行った．被験者は人物，娯楽，ビジネスの各ジャンルで興味のあるキーワードをアンケートに記入した．その後，提案システムを使用して，記入したキーワードについて検索を行った．なお，利用者の各ジャンルの検索順序や検索に費やす時間は指定していない．各ジャンルの検索結果の参照が終了した時点で，被験者はシステム評価に関するアンケートに回答した．アンケートの項目は，被験者の年齢と性別，三つのジャンルについて検索してどう感じたかを3段階評価，提案システムを使用したい度合いについて4段階評価で構成した．また自由記述欄を設け，上記質問に対する回答の理由や使用したい場面について調査した．

表2, 3に検索ジャンル別の満足度の調査結果，今後使用したいかの度合いの調査結果を示す．表2における人物についての検索では17名がA: 一つ以上の発見があったと回答し，4名がC: 意味の分からないものだったと回答した．Aと回答した被験者からは，「思わぬ人のつながりがあったから」や「意外な名前が検索結果に出たため」という意見があり，検索における有益な情報を得られていると確認された．また娯楽については18名がAと回答し，ビジネスについての検索でも16名がAと回答した．Aと回答した被験者からは，「自分の好きな物でも知らないことが多い」，「株主繋がりでの情報が見れて勉強になる」等，新たな発見をしたという意見が得られた．またB: 知っている内容だったと回答した被験者からは「知っている関連事項が出たから」，「予測出来る内容だった」など，検索結果に対して既知であるという意見が得られた．またCと回答した被験者からは「検索結果が多すぎて分からない」「検索結果が表示されなかった」等，情報量の多さから来る不満が得られた．

今後提案システムを使用したいかについては19名から肯定的回答が得られ「文字ばかりでなく，視覚的にわかりやすいため」や「知らない関係性がみえてきて面白い」などの意見が得られた．また，5名から否定的回答が得られ「情報量が極端だから」や「どのようなワードでもヒットする検索結果であれば使用したい」等の意見が得られた．

表 2 想定外検索のジャンル満足度調査 (括弧内は%を示す)

ジャンル	A	B	C
人物	17(65%)	5(19%)	4(16%)
娯楽	18(68%)	4(16%)	4(16%)
ビジネス	16(62%)	2(7%)	8(31%)

A: 一つ以上の発見があった
B: 知っている内容だった
C: 意味の分からないものだった

表 3 今後提案システムを使用したいかの調査

	A	B	C	D
度数(%)	1(4%)	18(68%)	7(28%)	0(%)

A: とても使用したい　B: 使用したい
C: 使用したくない　D: とても使用したくない

3.4 考察

前節の評価実験について考察する．ジャンル「人物」および「娯楽」において，分野ごとに項目を分類して，表2で一つ以上の発見があったと回答

表4「人物」の項目分けにおける「使用したいか」の分析(一つ以上の発見があったという回答のみ)

	歌手	俳優	その他
A	0	0	0
B	4	6	2
C	6	1	1
D	0	0	0
小計	10	7	3

A: とても使用したい　B: 使用したい
C: 使用したくない　D: とても使用したくない

表5「娯楽」の項目分けにおける「使用したいか」の分析(一つ以上の発見があったという回答のみ)

	漫画・アニメ	スポーツ	その他
A	0	0	0
B	2	3	4
C	4	0	0
D	0	0	0
小計	6	3	4

A: とても使用したい　B: 使用したい
C: 使用したくない　D: とても使用したくない

した被験者に対して提案システムを使用したいかについて調査した．表4, 5に「人物」と「娯楽」についての調査結果を示す．ここで歌手であり俳優である「福山雅治」のような場合では，それぞれの項目で数を数えた．また項目数が2件以下の場合はその他として分類した．表4の歌手の結果では肯定的回答と否定的回答がほぼ同数であったが，俳優に関しては肯定的意見が多かった．歌手を選んだ被験者からは,「検索結果が多いため，つながりがわかりにくい」という否定的な自由回答が得られた．一方で別の被験者における俳優については「知っている情報が多く表示されていたが，キーワードと同じジャンルの情報が見れたり、新しい発見があったため」と肯定的な回答が得られた．他の被験者でも同様の傾向が確認された．

表5の娯楽の場合では，漫画・アニメに関しては肯定的回答と否定的回答はほぼ同数であるが，スポーツでは肯定的意見のみが得られた．スポーツを検索した被験者は，そのスポーツに関する経験者であるため，検索内容に関して親近感があったと考えられる．自由記述においても同じ身長の選手を見つけて発見したという回答が得られた．

また，提案システムをどのような場面で使用したいかという自由記述では以下の回答が得られた．

- よく知る人物や物などの共通点を調べたい時
- キーワードと直接関係していない情報(例えば同一ジャンル)を発見するときに使用したい
- 就活などで興味のある会社に関連した会社が知りたいとき

上記の自由記述のように被験者から検索キーワードと関連するリソースを取得したいという要求があり，提案システムの目的と合致していることが確認された．現行システムではWikipediaのデータ構成上の理由で人物のマイニング精度のみが高いため，他の検索項目に対して精度を向上させることが今後の検討課題として挙げられる．また自由記述で得られた同一ジャンルの検索や企業情報のマイニング方法の検討を行う予定である．

4. おわりに

本稿では，利用者の新たな発見のために，Wikipediaを用いた想定外検索システムの提案及び実験を行った．実験の結果，検索キーワードのリソースから知識階層を抽出し，抽出語句を検索キーワードとした出力結果が確認された．評価実験の結果，提案システムで検索された出力結果から「調べてどう感じましたか」という質問に対して「一つ以上の発見があった」という回答が65%得られた．また「今後使用したいか」については72%の肯定的回答が得られ，提案システムの有効性が確認された．一方で，利用者の検索キーワードの興味の度合いによっては，利用者に対して有益な情報が提示できないという回答が得られた．

今後の課題として，実験結果のリソース数によって，提示する情報量を動的に変化させるグラフ提示が挙げられる．現行システムにおけるリソース過多による視認性の低下を改善する．また，SNSなどの人脈やブログ，同一業種企業，アニメや声優など，特定の分野に特化したリソースを提供することにより，利用者の興味を持つ分野における想定外検索を提供する．利用者の興味の度合いが高い対象を検索キーワードとすることで，想定外検索としての有用性の向上が期待される．

参考文献

[1] International Data Corporation: http://www.idc.com/
[2] 喜連川優: “情報爆発のこれまでとこれから”, 信学会誌, Vol. 94, No. 8, pp. 662–666. (2011)
[3] Google: https://www.google.co.jp/
[4] V. Gupta and G. S. Lehal: “A Survey of Text Mining Techniques and Applications”, *J. of Emerging Technologies in Web Intelligence*, Vol. 1, No. 1, (Aug. 2009)
[5] 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道: “文書クラスタリングによるトピック抽出および課題発見”, 社会技術研究論文集, Vol. 5, pp. 216–226 (2008)
[6] reflexa: http://labs.preferred.jp/reflexa/
[7] kizasi http://kizasi.jp/
[8] 中山浩太郎, 伊藤雅弘, E. Maike, 白川真澄, 道下智之, 原隆浩, 西尾章治郎: “Wikipedia研究のサーベイ”, 情処論データベース, Vol. 2, No. 4, pp. 49–60 (2009)