

RDF に対するハイブリッド Web マイニング

中山浩太郎 † 原 隆 浩† 西尾 章治郎†

セマンティック Web マイニングにおいては、RDF/OWL により Web リソース間の関係が定義され、「よく」構造化されているデータが十分に存在していることが前提条件となる。しかし、RSS など実運用されているセマンティック Web のデータは、いまだその多くがテキスト混じりの非常に浅いリンク構造であり、半構造化データの領域を超えていないのが実情である。そこで、本研究では、RDF の半構造化部分と構造化部分を並行してマイニングする手法である「ハイブリッド RDF マイニング」を提案する。本手法により、RSS のように比較的構造化されていない RDF に対するセマンティック Web マイニングを実現する。
キーワード セマンティック Web, RDF, Web マイニング

Hybrid Web Mining for RDF

KOTARO NAKAYAMA † TAKAHIRO HARA †
and SHOJIRO NISHIO†

In order to improve Semantic Web Mining, as a precondition, there have to be enough data which are “well”-structured by linking to other web resources. However, Semantic Web data in real world, such as RSS, are just semi-structured documents in most cases, because the main part of content is still mixed with text data. In this paper, we propose “Hybrid RDF Mining,” a new method to mine the structured and semi-structured part in RDF at the same time. Our approach accomplished Semantic Web Mining for semi-structured data such as RSS.

Keywords Semantic Web, RDF, Web Mining

1. ま え が き

近年、インターネットのユーザ数は爆発的に増加し、今もなお増加の途をたどっている。インターネットは既に情報通信の重要なインフラとしての位置を確立し、生活やビジネスになくってはならない存在となっている。この結果、様々な企業・個人・組織がコンテンツを提供しており、その総ページ数はもはや予測不能な領域にまで達している。このような背景の下、Web を文書のデータベース（コーパス）と見立て、膨大な量の情報から有益なデータを抽出する Web マイニングに関する研究が注目を集めている。Web マイニングと一言で言っても、その研究領域は非常に幅広く、コンテンツ（HTML 等）の内容を解析する自然言語処

理に近いものや Web リソース間（RDF）の関係を解析するもの、ユーザの行動履歴を分析するものなど、データの種類・解析技術共に多種多様である。Web マイニングは、膨大なコンテンツを持つ Web のポテンシャルを利用しようという目標の下、データベース・自然言語処理・情報検索・データマイニングなどさまざまな側面から研究が進められている。

このように注目されている Web マイニングであるが、近年、セマンティック Web²⁾ の登場により新たな局面を迎えている。それが、セマンティック Web マイニングである。セマンティック Web マイニングでは、セマンティック Web のデータフォーマット（RDF/OWL）によって定義されたデータをマイニングするため、Web リソース間の関係が「よく」構造化されているデータが豊富に存在していることが前提条件である。しかし、RSS など実運用されているセマンティック Web のアプリケーションの中身は、`rss:title` や `rss:describe` などの部分に見られるように、テキスト混じりの非

† 大阪大学大学院情報科学研究科マルチメディア工學専攻
Department of Multimedia Engineering, Graduate School
of Information Science and Technology,
Osaka University

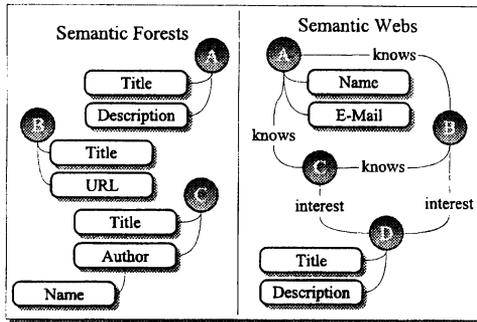


図 1 Semantic Forests と Semantic Webs

常に浅いリンク構造がそのほとんどであり、半構造化データの領域を超えていないのが実情である。これは、一般に「Semantic Forests」⁶⁾と呼ばれ、小さな Web リソースのツリーがいくつも分散して存在している状態である。セマンティック Web が目指す「よく」構造化された状態ではなく、Web リソース同士が互いに大きなグラフを構成する「Semantic Webs」をいかに実現するかが課題となっている。図 1 に Semantic Forests と Semantic Webs の概念を示す。

Semantic Forests の状態のデータに対しては、Web リソース間の関係を解析するセマンティック Web マイニングの技術を適用できないため、この解決が主要な課題の一つとなっている。そこで、本研究では、セマンティック Web において、半構造化部分と構造化部分を並行して解析することで、知識獲得を目指す手法「ハイブリッド RDF マイニング」を提案する。本手法では、まず RSS のように、浅いリンクで構成されている Web リソースに対し、テキストマイニングを適用することにより、他の Web リソースへの参照を生成する。これにより、Semantic Forests 状態のデータを Semantic Webs の状態に変化させる。そして、これら生成されたデータと既存の RDF/OWL データの参照関係を解析することで、セマンティック Web マイニングによる概念辞書の構築を行う。本手法は、RSS に代表されるような浅いリンクにより構成されているデータについても、適用できるのが特徴である。本論文では、本手法についてその詳細なアルゴリズムと、実装の状況について述べる。

本稿の以下では、2 章で本研究に深く関わるセマンティック Web マイニングについて述べ、3 章で本手法の詳細について記述する。また、4 章では本手法を元に開発したシステムについて述べる。最後に、5 章でまとめと今後の展開を記述する。

2. セマンティック Web マイニング

本章はまず、Web マイニングとセマンティック Web マイニングの現状を俯瞰することで、本研究のアプローチとその意義を明確にする。

2.1 Web マイニング

Web マイニングは情報を抽出する対象のデータの視点から、「Web structure (構造) マイニング」「Web usage (利用) マイニング」「Web content (内容) マイニング」の 3 つに分類されるのが一般的である⁵⁾。

Web 構造マイニングは、Web サイトの構造や Web ページ間の関係を解析する手法である。Web 構造マイニングでは、Web サイトの構造やハイパーリンク構造を解析することで、ページ間の影響度や類似度を計算することを目的としている。リンクベースのページ分類や、Google で活用されている PageRank のアルゴリズム⁸⁾などが Web 構造マイニングの代表例である。PageRank では、他のページから参照されている回数をカウントすることで各ページの重要度を算出し、検索エンジンに利用している。

Web 利用マイニングは、利用者の行動履歴など、利用ログを解析する手法である。Web 利用マイニングでは、サーバサイドに蓄積された利用ログなどをマイニングすることで、サイトの利用者傾向を調査することやユーザビリティ検証、ボトルネックなどを発見することを目的としている。現在の Web マイニング研究の多くは、このユーザの行動ログを解析する Web 利用マイニングであるといわれており、その有用性から、企業の研究者も数多く参入している。

Web 内容マイニングは、Web ページの内容 (コンテンツ) を解析する手法である。Web 内容マイニングでは、ページの内容を解析することで、重要単語やページの構造などの情報を抽出し、ページのカテゴリや要約などを行うことを目的としている。例えば、ニュースサイトの記事に含まれる単語同士の共起性の発見は、最も代表的な例の一つである。また、テキストだけでなく、音声、ビデオ、メタデータなども Web コンテンツに分類され、これらのハイパーメディアを対象とした研究も盛んに進められている。しかし、(ハイパー) テキストが研究の主流であることは変わらず、半構造化・構造化されたデータである HTML や XML が主な研究対象となっている。これらのデータは、さらに 1) 自由記述されたテキスト、2) 半構造化データ、3) 構造化データの 3 つに分類することができる⁷⁾。

2.2 セマンティック Web マイニング

このような状況の下、最近新しい Web マイニングの研究領域であるセマンティック Web マイニングが注目されている。セマンティック Web マイニングは、二つの研究領域「セマンティック Web」と「Web マイニング」を統合するものである¹⁾。その主な目的は、構造化されたデータ（セマンティック Web のデータ）を利用することで、Web マイニングの精度を向上させることにある。また、Web マイニングの技術がセマンティック Web のデータ（RDF や OWL など）を構築する目的で利用されることも含めてセマンティック Web マイニングと呼ばれている。セマンティック Web では、URI を利用して Web リソース同士の関係を定義したものが主なコンテンツとなるため、コンテンツの中身を解析するということは、Web リソース同士の構造を解析することとほぼ同義といえる。つまり、セマンティック Web マイニングでは Web 内容マイニングと Web 構造マイニングが統合され、その区別がほぼなくなる。

具体的な研究例としては、Maedche らはセマンティック Web のデータをマイニングすることで、オントロジの構築を試みる研究⁹⁾を進めている。Maedche らの研究の目的は、Web マイニングの技術を利用することでオントロジや概念を（半）自動的に生成することである。この研究は、非常にチャレンジングな取り組みであり、機械学習・情報検索・エージェントなどの技術を利用してセマンティクスを発見しようとしている。

3. ハイブリッド RDF マイニング

セマンティック Web の基盤データフォーマットである RDF は、従来の HTML などと比べるとデータの構造化が高度に洗練されており、Web リソース間の様々な関係を厳密に定義することができる。セマンティック Web マイニングを行うためには、前提条件として、「よく」構造化されたセマンティック Web のデータが必要となる。しかし、現在最も利用されている RSS や FOAF では、Description 要素などをはじめとするデータの多くは未だに単なるテキストもしくは HTML で記載されている場合が多く、依然 Semantic Forests の状態であることがわかる。当然、foaf:knows など、リソース間の関係が定義されているエンティティも存在するが、Web リソースのコーパスとして利用できるほど十分なデータ数と種類が揃っていないのが実情である。これは、RDF を自動生成するためのエンドユーザ用アプリケーションが未だ開発途上にあり、操

作性・有用性において実用に耐えうるソフトウェアが存在しないことに起因する。長期的に見れば、これらの問題は徐々に解決され、RDF によりリソース間の関係が厳密に定義される方向にあるが、短期的、少なくともここ数年では、現在の RSS や FOAF データのようなテキストによる記述の混在するデータが主流となることが予想される。

筆者らが提案する RDF に対するハイブリッド Web マイニングは、このようなあまり構造化されていないセマンティック Web データに対する解析手法である。本章では、その詳細について解説する。

3.1 パーソナルオントロジ

機械理解可能なデータフォーマットにより記述される Sematic Web においては、人間の変わりにエージェントがそのフロントエンドとなり、情報を収集・蓄積・処理・発信することが期待される。しかし、Web のような分散環境の中では、ユーザ同士の意見の相違などによる「矛盾」や、参照がたどれない「不完全な知識」といった問題が発生する。また、個人の属する文化や組織によって、概念構造は異なるはずであり、さらには、同じ組織に属していても、個人によって概念構造は異なるはずである。このような状況の中で、ロバストな知識体系を維持するためには、各ユーザのフロントエンドとして、エージェントがユーザへパーソナライズ（Web の巡回・環境適応・知識の取捨選択）した概念辞書を持ち、メンテナンスするというアーキテクチャが自然である。さらに言えば、ユーザからのクエリ処理という側面から見ても、クエリが発せられる度にリソース間の関係をたどって Web を巡回したのでは、現実的な処理スピードは望めない。そこで、筆者らはこれらの問題を解決する手法として、「パーソナルオントロジ」（Personal Ontology）を提案する。パーソナルオントロジは、各ユーザもしくは各組織が保有する概念体系辞書であり、各概念間の関係に信頼度係数（CF）を設けることで、各ユーザに適応することが可能になっている点が大きな特徴である。これは、二つの異なる（矛盾）する概念が Web 上に存在したときに、正か誤かを判断するのではなく、パラメータを変化させることによって、エージェントが環境（ユーザ）に柔軟に適応していくことを目的としている。パーソナルオントロジの概念を図 2 に示す。

この例は、単語「Hospital」に関連する語とその関係を示したものである。例えば、Hospital は、Medical Institute の一種であり、Surgery Medicine や Internal Medicine といった部署を持つといった概念が記述されている。これらの語彙同士は、信頼度係数が設け

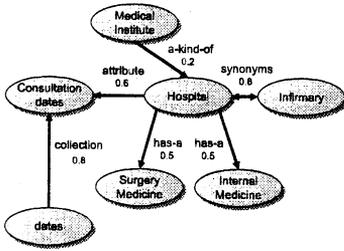


図 2 パーソナルオントロジの概念

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
  xmlns="http://purl.org/rss/1.0/"
  xmlns:rdf="http://www.w3.org/rdf#"
  xml:lang="ja">
  .
  .
  <item rdf:about="http://hoge.mac/tiger">
    <title>Tiger</title>
    <link>http://www.apple.com/tiger/</link>
    <description>
      Tiger is really cool. I like this OS.
    </description>
  </item>
  .
  .
</rdf:RDF>
```

図 3 RSS の例

られた関係 (a-kind-of など) によって相互にリンクされる。提案手法では、Word Net¹⁰⁾ のデータをパーソナルオントロジの初期データとして利用し、運用中はエージェントを介して概念体系の更新を行う。

3.2 RDF に対する Web リソースマッピング

RSS は、セマンティック Web 技術の中でも、FOAF と並び最もよく利用されている RDF のアプリケーションの一つである。しかし、その内容を見てみると、URI と RDF スキーマによる Web リソース間の関係が定義されているケースは極めて少ない。図 3 は、Weblog で利用されている RSS の例であるが、タイトルやエントリの説明などのテキストデータが中心であり、いまだ Semantic Web の目指す完全構造化されたデータには程遠いことがわかる。また、Web リソースとの関係を定義するための唯一の属性である `rdf:about` 属性も、この属性を誰がどのように定義するのかという問題が残ったままである。

そこで、本研究では、RSS に対する Web リソースマッピングアルゴリズムを提案する。これは、RSS の内容を解析し、関係の深い Web リソースを探索することで、RDF:about 属性を自動的に付与するものであ

る。本アルゴリズムのフローとその詳細を以下に示す。

3.2.1 RDF データの特徴分析

RDF のテキスト部分を抽出し、単語リストを作成する。各単語に対し、プリプロセッシングとして、Stemming 処理と Tagger によるタグ付けを行い、名詞部と未知語の抽出を行う。

3.2.2 Lead 法による重要度算定

重要語の抽出方法としては、tfidf 法をはじめとして語の出現頻度から統計的な自然言語処理で重要度を算出する方法が一般に利用される。しかし、Weblog のように、比較的短い文章が分散的に存在するような状況では、出現頻度をカウントすることが困難である。また、tdidf 法では、文章中には出現しないが、暗にその事象を示唆しているような文章からは重要語を抽出できないという問題もある。そこで、本研究では、段落や文の先頭にくる語ほど重要であるという考えに基づく Lead 法⁴⁾ をベースにした重要度算出を行い、語と語の関係を考慮してさらに重要度計算をするという二段階の重要度算出アルゴリズムを提案する。このとき、単語 w とその重要度 p の組みのリストを W とし、以下のように定義する。

$$W = \{(w_1, p_1), (w_2, p_2), \dots, (w_i, p_i), \dots\}$$

重要度 p_i は、以下の式で算出した (l は単語 w_i の出現回数)。

$$p_i = \sum_{k=1}^l (\text{length}(\text{doc}) - \text{position}(w_i, k)) / \text{length}(\text{doc})$$

$\text{length}(\text{doc})$ は段落 (ドキュメント) に含まれる単語の総数、 $\text{position}(w_i, k)$ は k 回目出現した単語 w_i の出現場所とする。また、RSS であれば、`title` 属性、FOAF であれば `interest` 属性など、各エントリの中でも重要なエンティティに対しては、重要度を高く設定する目的のために、 p_i に定数 α を乗じた値を重要度とする。

例えば、以下のような RSS エントリについて考察する。

```
<item>
  <title>Thinkpad</title>
  <description>
    My new X41 arrived to my home a few minute ago...
  </description>
  .
  .
</item>
```

表 1 概念間関係を考慮したの係数

関係	$cf(r)$	例
A same-as B	1.00	PC はパソコンと同義だ
A is-a-kind-of B	0.75	Thinkpad(A) は PC(B) の一種
B is-a-kind-of A	0.50	Thinkpad(B) は PC(A) の一種
A relates-to B	0.25	Thinkpad(A) は PC(B) に関係している

上記のようなエントリからは、以下のような重要語とその重要度のリスト W が抽出できる。

$$W = \{(Thinkpad, 1.0), (X41, 0.8), (home, 0.7), \dots\}$$

3.2.3 Web リソース関係を考慮した重要度算定

上記の手順により抽出された重要語に対し、第二ステップとして単語間の関係を利用して更に重要度の解析を行う。具体的には、単語が持つほかの語との関係をパーソナルオントロジから抽出し、再帰的に探索していく中で、各単語の重要度をスコアリングしていく。このとき、各単語に關係する単語リストを取得する関数を $GetRelations(w)$ と定義する。単語リストは、語 w に關係する単語を w_j 、關係の種類を r_j 、關係に付与されている確信度を c_j とし、次式 R_w により定義する。

$$R_w = \{(w_1, r_1, c_1), (w_2, r_2, c_2), \dots, (w_j, r_j, c_j, \dots)\}$$

前述の Lead 法によって得られたすべての重要語 w_i に対し、以下の再帰アルゴリズム RE を適用する。

Algorithm $RE(w, p)$

- 1 if $i < \delta$ then return;
- 2 $R_w = GetRelations(w)$;
- 3 for each $(w_j, r_j, c_j) \in R_w$ do
- 4 $s = Score(p, c_j, r_j)$;
- 5 $S_w = S_w + s$;
- 6 $RE(w_j, s)$;

ここで、スコアリング関数 $Score$ は次式のように定義する。

$$Score(i, c, r) = i \cdot c \cdot cf(r)$$

$cf(r)$ は、各 Web リソース間の關係の種類 r に基づいた探索の重み付け係数を求める関数である。 $cf(r)$ の係数一覧表を表 1 に示す。

本アルゴリズムでは、探索が枝別れるするに従い、重要度が低下していく一方で、Lead 法によって発見し

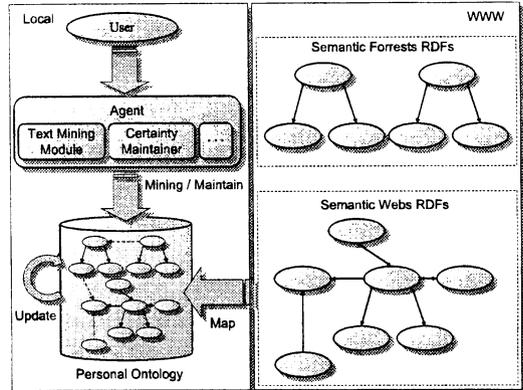


図 4 システム全体図

た重要語とより多く關係する単語についてはそのスコアが高くなるように設計した。最後に、単語 w のスコア合計値 S_w について、降順にソートし、上位 n 件を抽出することで、そのエントリに対する Web リソースの概念マップを生成する。この概念マップは、実際の RDF に直接埋め込まれるわけではなく、その写像をパーソナルオントロジに蓄えておくことで、次のマイニングに利用する。

3.2.4 パーソナルオントロジの更新

パーソナルオントロジは、ユーザが所有する RDF データが変更されたとき、前節で解説した Web リソース関係を考慮した重要度算定が行われたときの二つのタイミングで更新される。Web リソース関係を考慮した重要度算定が行われたときには、単語 w のスコア合計値 S_w の上位 n 件の単語について、単語間の信頼度係数の値を上昇させる。言い換えれば、Web リソース同士の共起性に基づき關係係数を更新しており、よく関連する概念同士の信頼係数がより向上するように設計している。これは、語の共起性に基づく情報検索¹¹⁾ の考え方を Web リソースに適用したものである。また、単語同士の關係が定義されていない場合は、新たに關係「A relates-to B」を生成する。この更新作業により、パーソナルオントロジは常にパーソナライズされた状態を保つ。

4. 実装

本研究では、前章の提案手法に基づき、パーソナルオントロジを利用した RSS への自動アノテーションシステムを構築した。図 4 にシステム構成を示す。

パーソナライズオントロジの初期化の手順としては、まずローカルマシン上に仮想の Web サーバを構築し、

次に WordNet のデータを RDFS/OWL フォーマットに変換・公開し、最後にシステムに読み込む。このとき、データフォーマットは、基本的には既存研究である文献³⁾の名前空間を踏襲した。しかし、この研究では Web リソースの関係は定義しているものの、その意味データに対するラベル(表層ラベル)に関して考慮されていないため、自然言語とのマッピングができないという問題があった。そこで、新たに `rdfs:label` 属性によりメタデータを付与することで、この問題を解決した。このラベルデータは、テキストマイニングから RDF へマッピングする際に必要となる。語彙「hospital」に関する RDFS/OWL のサンプルを以下に示す。

```

...
<wn:Noun rdf:about="&wn;106690409" />
<wn:WordObject rdf:about="&wn;hospital" />
<rdf:Description rdf:about="&wn;106690409">
  <wn:wordForm rdf:resource="&wn;hospital" />
  <rdfs:label>hospital</rdfs:label>
  <wn:glossaryEntry>
    a medical institution where sick or
    injured people are given medical or ...
  </wn:glossaryEntry>
  <wn:hyponymOf
    rdf:resource="&wn;106689915" />
</rdf:Description>
...
<rdf:Description rdf:about="&wn;106689915">
  <wn:wordForm
    rdf:resource="&wn;medical_institution" />
  <rdfs:label>medical institution</rdfs:label>
</rdf:Description>
...

```

106690409 や 106689915 といった数字は、WordNet の中で各語彙に与えられている連番であるが、本研究では語彙を一意に特定するための URI として利用している。

実装したシステムにより、RDF により定義された Web リソースの写像をパーソナルオントロジに記録しておき、さらに構造化されていない箇所に関しては、Web リソース関係を考慮した重要度算定により概念マップを生成することが可能となる。その結果、本稿の冒頭で述べた Semantic Forests の問題を解決することができる。

5. まとめと今後の展開

本研究では、RSS や FOAF など、現実に利用されているセマンティック Web データの中で、構造化されていない箇所に対してテキストマイニングによる意味付けを行い、パーソナルオントロジに写像することで、セマンティック Web マイニングが可能になるハイブリッド Web マイニングを提案した。今後は、実験用のサイト(セマンティック Web ポータル)を構築し、RDF データの収集と評価実験を行う予定である。

また、課題としては、テキストマイニングの精度向上と、メタデータ記述インタフェースの実現が挙げられる。テキストマイニング精度の向上としては、まだまだ改良点が残されており、例えば *cf* 関数のパラメータ調整や連語への対応などが挙げられる。連語処理に対応するためには、N-Gram を利用した解析を導入する予定である。特に、各ドメイン・カテゴリに分類した N-Gram 解析を行うことで、新しいテクニカルタームや語の置き換えなどを含むデータにも対応できると考えられる。また、メタデータ記述のためのインタフェースとしては、Weblog ベースのエージェントインタフェースが必要になると考えている。これは、セマンティック Web の問題として、メタデータ不足が深刻な問題として叫ばれているが、その有力なブレークスルーとして Weblog に注目が集まっているためである。Weblog の利便性と、トラックバックによる相互接続性は、セマンティック Web のメタデータ不足を解決しうる可能性を秘めており、本研究においてもセマンティック Weblog という視点から、対応を進めたいと考えている。

参 考 文 献

- 1) B. Berendt, A. Hotho, and G. Stumme. "Towards Semantic Web Mining," Proc. of the First International Semantic Web Conference, pp. 264-278, 2002.
- 2) T. Berners-Lee, J. Hendler, and O. Lassila. "The Semantic Web," Scientific American, pp. 35-43, 2001.
- 3) C. Ciorascu, I. Ciorascu, and K. Stoffel. "knOWler Ontological Support for Information Retrieval Systems," Proc. of SIGIR 2003 Conference, 2003.
- 4) H. P. Edmundson. "New Methods in Automatic Extracting," Journal of ACM, Vol. 16, No. 2, pp. 264-285, 1969.
- 5) F. M. Facca and P. L. Lanzi. "Mining Interesting Knowledge from Weblogs: A Survey," Data

- and Knowledge Engineering, Vol. 53, Issue 3, pp. 225-241, 2005.
- 6) G. A. Grimnes, P. Edwards, and A. Preece "Learning Meta-descriptions of the FOAF Network," Proc. of International Semantic Web Conference 2004, pp. 152-165, 2004.
 - 7) R. Kosala, H. Blockeel, and K. Leuven. "Web Mining Research: A Survey," ACM SIGKDD Explorations, Vol. 2, Issue 1, pp. 1-15, 2000.
 - 8) P. Lawrence, B. Sergey, M. Rajeev, and W. Terry "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report, Stanford Digital Library Technologies Project, 1999.
 - 9) A. Maedche and S. Staab, "Ontology Learning for the Semantic Web," IEEE Intelligent Systems, Vol. 16, Issue 2, pp. 72-79, 1999.
 - 10) G. A. Miller, "WordNet: A Lexical Database for English," Communications of the ACM, Vol. 38, No. 11, pp. 39-41, 1995.
 - 11) H. Schutze and Jan O. Pedersen, "A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval," International Journal of Information Processing and Management, Vol. 33, Issue 3, pp. 307-318, 1997.