1D2-5

日本語 Wikipedia マイニングと Folksonomy タグに基づく 領域オントロジー構築支援

Extending DODDLE-OWL with Wikipedia Mining and Folksonomy Tags

手島拓也[*1] Takuya Tejima　森田武史[*1] Takeshi Morita　和泉憲明[*2] Noriaki Izumi　山口高平[*1] Takahira Yamaguchi

[*1] 慶應義塾大学 Keio University

[*2] 独立行政法人 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

The user-generated and semi-structured information resource is widespread on the emergence of Web2.0. In this paper, we propose a way to construct domain ontology with a new body of knowledge created by the resource of wikipedia and folksonomy tags. Therefore, it means to solve the problem of a current DODDLE-OWL.

1. はじめに

領域オントロジー構築支援ツールとして我々が開発した DODDLE-OWL[Morita 06]は，階層関係定義に汎用オントロジーを参照し，非階層関係定義には専門文書等のテキスト中に出現する語の共起関係を得ることで，人手によるオントロジー構築のコストを低減させることが可能である．しかし，何も構造を持たないフリーテキストをリソースとしている故に，概念定義の際に出現するゴミは多く，依然，人手による構築コストは大きい．一方，現在の Web2.0 というユーザ参加型の大規模なネットワーク構造では，フリーテキストとは異なるハイパーリンクやフィードを活用した半構造化情報資源が広がりをみせている．中でも情報鮮度・語彙網羅性の優れた百科事典 Wikipedia と，大衆により付加された情報である Folksonomy タグがその代表例である．

本研究は，そのような半構造化された情報資源である Wikipedia と Folksonomy タグを利用した領域オントロジー構築支援を目的とする．ただ，それらの情報資源はゴミが多く完全に構造化されているわけではないため，オントロジーへ直接結び付けることは難しい．そこで我々は，Wikipedia に対するリンクマイニングを行って得られた語彙間関連度をリソースとした関連概念抽出手法を提案する．また，外部からの付加的な情報資源を概念定義に組み込む第一の試みとして Folksonomy タグの利用を検討する．それらは同時に，最新語彙等の未知語の定義が困難であり，テキストに出現しない語の定義が不可能である現状の DODDLE-OWL の持つ問題点を解消する可能性を持つ．ケーススタディとして，ある IT 関連の文書を対象に Wikipedia と Folksonomy タグに用いられる語彙に Wikipedia マイニングに基づいて得られた語彙間の関連度を用いて概念定義を行う場合と，従来までの語の共起性による手法との比較を行い，その評価について述べる．

2. 領域オントロジー構築支援

本研究では，対象文書に対して「コンテンツ内容を表すキーワード」と「コンテンツ内容を表すタグ」をリソースとして用いた領域オントロジー構築支援を以下の①～③の流れで行う．

①ドキュメント解析によるコンテンツ内容を表すキーワード抽出

②ドキュメントと関連性の高いタグの抽出

③抽出したタグ・キーワードをリソースとした概念関係定義

上記の①～③に関する詳細な計算方法等はそれぞれ 4.2 節～4.4 節に記す．いずれも 3.1 節で示す日本語 Wikipedia をリソースとした語彙間の関連度計算結果を用いており，対象文書の内容に深く関係したキーワード・タグの抽出，概念関係の抽出を行う．下図1がその全体のイメージである．

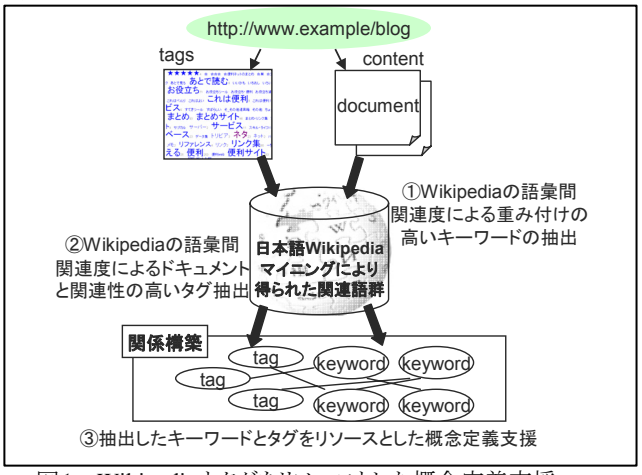

図1　Wikipedia とタグをリソースとした概念定義支援

3. リンクマイニング

リンクマイニングとは，近年のユーザ参加型 Web によって形作られた大規模なネットワーク構造に対する新しい解析手法であり，Getoor[Getoor 05]らの調査によれば，ノードに対するランキング，分類，クラスタリング等を目的としたものである． 特に，密なリンク構造を持つ Wikipedia に関してリンクマイニングを行い，有益な情報を抽出する手法である「Wikipedia マイニング」を中山ら[中山 06]が提案している．Wikipedia マイニングは，①最新の語や概念への対応，②自然言語処理による形態素への分割や同義語・多義語処理等の語彙間の関連性を解析する前段階処理での精度低下が生じない，という特徴を持つ．

3.1 リンクマイニングによる語彙間関連度計算

本研究では語彙間の関連度を計算する際，中山らの Wikipedia マイニングの技術を用いる．以下，その概要を説明する．Wikipedia におけるすべての Web ページ（記事）の集合

連絡先：手島拓也，山口高平，慶應義塾大学理工学部
〒223-8522 神奈川県横浜市港北区日吉 3-14-1
Tel: 045-566-1614，E-mail: {t_tejima,yamaguti}@ae.keio.ac.jp

を $P=\{p_1, p_2, p_3, ..., p_n\}$ と定義する．このとき，ページ p_i $(1 \le i \le n)$は，Forward Link と Backward Link の2種類のリンクを持つ．p_i の Forward Link は，ページ p_i から別のページへジャンプするリンクの集合であり，$F_{p_i}=\{f_{i1}, f_{i2}, f_{i3}, ..., f_{in}\}$ と定義する．また，Backward Link は別のページから p_i へジャンプするリンクの集合であり，$B_{p_i}=\{b_{i1}, b_{i2}, b_{i3}, ..., b_{il}\}$ と定義する．また，ある記事 p_i が参照されたときに別の語彙（記事）へ転送するリンク（リダイレクトリンク）の集合を $R_{p_i}=\{r_{i1}, r_{i2}, r_{i3}, ..., b_{ik}\}$ と定義する．p_i に関する語彙の一覧とその関係の強さを求める再起探索アルゴリズム *RE* を以下のとおりに定義する．

```
Algorithm RE(pi,weight,depth)      8  for each (pj) ∈  Bpi do
1  if depth > n then return;       9    score = weight/|Bpi|;
2   Fpi = GetFowardLinks(pi);      10   Spj = Spj + score;
3  for each (pj) ∈  Fpi do         11   RE(pj,score,depth+1);
4    score = weight/|Fpi|;         12  Rpi = GetRedirectLinks(pi);
5    Spj = Spj + score;            13 for each (pj) ∈  Rpi do
6    RE(pj,score,depth+1);         14   RE(pj,weight,depth);
7   Bpi = GetBackwardLinks(pi);
```

単に Forward Link と Backward Link の有無を解析するのではなく，その先のページを再帰的に探索することで，語どうしの関係の強さを計算する．このアルゴリズムによって，ページ p_i に対する関連度を関係度の高い順に抽出することができる．

4. タグ・キーワード間関連性評価

Folksonomy によりある対象（インスタンス）へ付与された情報は多量であり，それらユーザによって記述されるタグは対象と様々な関連性を持つ．対象がドキュメントタイプの場合にはその要旨を的確に表すものだけでなく，ドキュメント中に登場しない語彙（あるキーワードを抽象化したもの等）までも多く付与されているのが現状であり， 対象とタグとの関連性を明らかにすることは大衆の合意を得た概念定義に結びつく可能性がある．

ここで筆者は，ユーザがタグに付与するセマンティクスをある程度分類できることに注目し，以下①～⑧のように分類した．

①対象 URL のコンテンツ内容のキーワード・主旨を表すもの
②対象 URL のタイトル・ラベルを表すもの
③対象 URL の作成者を表すもの
④対象 URL とタグがインスタンスークラス関係であるもの
⑤対象 URL のコンテンツが持つ機能・サービスを表すもの
⑥対象 URL が作成された場所・国，または言語を表すもの
⑦ユーザによる対象 URL の評価を表すもの
⑧ユーザ自身の分類・整理のためのみに使われるもの

このようにタグは様々なセマンティクスを持つことがわかるが，ドキュメント中に登場する語との概念定義に利用することを念頭に置けば，①のコンテンツ内容を意味するタグが最も重要であるといえる．なぜなら，それらのタグはユーザが対象文書（URL）中に登場するキーワードに対し何らかの関連性を持って付与したものであり，概念定義に直接結びつく可能性が高いからである．

4.1 Folksonomy タグのクレンジング

Social Bookmark Service(以下 SBS)おいて付与されるタグの現状は非常に混沌としている．ここで言う混沌としているという意味は，「口語的なもの」，または「自分の分類のためのみに用いられる乱雑な語彙」などの電子辞書等を用いて機械が判断不可能な語彙の多さを表している．このような混沌としたタグを全て概念定義の対象語彙に含むことは効率ではない．したがって，本研究では日本語 Wikipedia によるフィルタリングを行う．

ここで，まず現状の Folksonomy タグに使用される語彙の混沌さの尺度を計るために，Wiki ベースの百科事典である Wikipedia（日本語）と電子辞書 EDR による語彙のマッチングを行う．実際に，日本の SBS において最も多くのユーザ数を持つ「はてなブックマーク」に注目し，付与されているタグを約 900,000 個収集し，重複を排除した後（約 28,000 個），約 18,000 個の日本語のタグのみを対象に「日本語 Wikipedia」と「EDR 日本電子化辞書」を用いてフィルタリングを行った結果が以下の表1である．

表 1 現状の Folksonomy タグのフィルタリング結果

対象日本語タグ数（重複無し）	日本語 Wikipedia とマッチする語彙数	EDR とマッチする語彙数
17,686	7,340	5,647

現状の日本語のタグとマッチする割合はそれぞれ日本語 Wikipedia が 41.5%，EDR が 31.9%であり，Wikipedia の方がタグに用いられる語彙が多いことがわかる．これは，既存電子辞書よりも，Web2.0 というユーザ参加型の Web（Wiki）において構築された「集合知」としての知識形態の方が，あらゆるユーザを取り込む Folksonomy タグに対して適応性が高いことを示している．

4.2 ドキュメント解析によるキーワードの重み付け

この節ではユーザのタグ付けによるセマンティクスを分類した際，「①対象 URL のコンテンツ内容のキーワード・主旨を表すもの」のタグを抽出するために日本語 Wikipedia マイニングによって得られた語彙間関連度を用いて，コンテンツと付与されたタグとの関連性評価を行う．

具体的な計算方法は，抽出した語彙について重複を排除した後，「あるキーワードについて，Wikipedia マイニングにより得られた関連語彙（上位最大 10,000 まで保存）に，解析対象のドキュメント中に用いられている他のキーワードが出現したらその関連度を追加」してゆき，総和を求める．したがって，ドキュメント中に存在するキーワード同士の関連性を計っていくことになり，他のキーワードとの関連性が高い語彙ほど総和が高くなり，低い語彙は総和が低くなる．

このような計算方法によってドキュメント中に出現するキーワードの重み付けがなされるため，上位のキーワードは対象ドキュメントのトピック（主旨）である可能性が高い．そして，これらのトピックと重みは対象ドキュメントと，そこに付与されたタグクラウドとの関連性を評価する際に用いる．また，この計算によって得られた結果は領域オントロジーを構築する際，ドキュメント中に出現するゴミとして扱われるクラス概念の排除が可能となる．

4.3 タグ・キーワード間の関連性評価

各タグについて日本語 Wikipedia マイニングによって得られた関連語彙と関連度を用いて，タグ・キーワード間の関連性を評価する．「タグ自身もしくはそのタグが持つ関連語彙がドキュメント中に存在するキーワードにマッチした場合，そのタグとキーワードとの関連度を加算していく」という手法を用いて計算する．重み付け結果が上位であるキーワードと多くヒットする関連語を持つタグが高いポイントを得ていくため，「①対象 URL のコンテンツ内容のキーワード・主旨を表すもの」というセマンティクスを持つタグが抽出可能となる．

4.4 抽出したタグ・キーワードの概念関係評価

対象文書から抽出したキーワード・タグそれぞれに日本語 Wikipedia マイニングにより計算された語彙間の関連度を持つため，どのキーワードが他のキーワードもしくはタグと関連性が深いかを計算することができる．具体的な計算手法は4.2 節のキーワードの重み付けで述べたものと同様であり，それぞれの抽出したキーワード・タグについて，Wikipedia マイニングにより得られた関連語彙に，他のキーワード・タグが出現したらその関連度をキーワード間，もしくはキーワード・タグ間の関連度として計算結果を求める．

Wikipedia マイニングに基づいて作成された語彙間の関連度は汎用的なものであるが，本研究においては文書中に登場する語彙との関連を計っているため領域オントロジーの構築が可能となる．また Folksonomy により付加されるタグはドキュメント中に存在する語彙（概念）以外の情報も数多く存在することに注目すると，概念定義対象のリソースとして用いることができ，ドキュメント解析のみでは得られない情報も得ることができる．

5. ケーススタディ

本ケーススタディの目的は，ドキュメント中に登場する要旨を表す語彙とそのドキュメントに付与されたFolksonomyタグの「コンテンツ内容を表すもの」について，日本語Wikipediaに基づいた語彙間関連度を用いた評価による，比較的最新の語彙を多く含む領域に対する非階層関係定義支援の有用性を確かめることである．「初心者がJavaを超高速で学ぶためのコツ」[*1]というIT関連の文書を対象とし，まず日本語形態素解析ツール chasen[chasen]を用いて文書と付与されたタグからWikipediaに掲載されている語を抽出する．それらをリソースとして本研究の手法により概念定義を行う場合と，従来までの語の共起性による手法との比較を行い，その評価について述べる．また，実験の考察について述べる．

5.1 日本語 Wikipedia によるタグクレンジング

対象文書に付与されているタグ全体に対して日本語 Wikipedia フィルタリングを行った結果，全 128 種類中 52 種類が残り，日本語のタグに関しては 83 種類中 45 種類が残った．これらを 4 章で示したタグのセマンティクス別に見ると，対象ドキュメントと関連度が高いと考えられる「①対象 URL のコンテンツ内容を表すもの」に関しては 23 種類中 19 種類の 82.6%が Wikipedia の掲載語とマッチし，対象ドキュメントと関連度が低いと考えられる「⑧ユーザ自身の分類・整理のためのみに使われるもの」に関しては，48 種類中 21 種類の 43.8%のタグがマッチした．したがって，Wikipedia によるタグクレンジングは概念定義の際に有益な①のタグは高い確率で残し，概念定義の際にゴミとして扱われる可能性が高い⑧のタグは比較的多く排除できることがわかる．

5.2 ドキュメント内のキーワード重み付け実験と結果

キーワードの重み付けを行う際に本研究では Wikipedia マイニングにより得られた結果を用いる．マイニングの手法は 3.2 節で示した方法を用いる．IT関連の語彙を中心に約 500 語を対象に関連度計算を行った．対象文書に 4.2 節で述べたコンテンツ解析によるキーワードの重み付け計算を行う．文書中に登場するキーワードは日本語 Wikipedia 辞書を追加した chasen を用いてキーワードを抽出し，重複を排除したところ 217 語得られた．各々のキーワードに対し他のキーワードとの関連度総和を求めた結果の上位5つが以下の表 2 である．

表 2 各キーワードの重み付け結果

1	クラス変数	2.008
2	集約	1.794
3	インスタンス	1.491
4	コンストラクタ	1.226
5	ポリモーフィズム	1.220

重みが高い語彙ほど，対象ドキュメントの内容・主旨を表している可能性が高い．

評価として，再現率（Recall）と適合率(Precision)を用いる．候補集合を「プログラムが対象ドキュメントのコンテンツ内容を表すキーワードであると予想した集合」，正解集合を「人間がコンテンツ内容を表すキーワードであると判断した集合」とした時の再現率と適合率のグラフは下図 3 に示す．

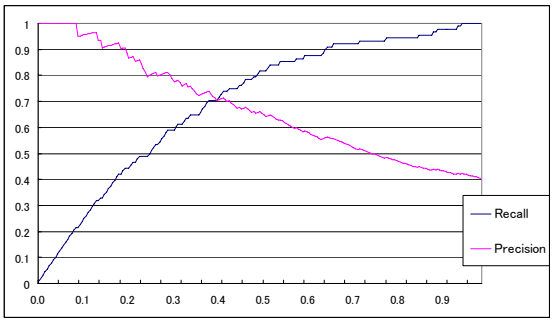

図 3 コンテンツ内容を表すキーワードの抽出結果

5.3 タグ・キーワード間の関連度計算の実験と結果

5.1 節で用いた対象 URL へ実際に付与されている 128 種のタグに対し，日本語 Wikipedia フィルタリングを行って得られた 53 語のタグと，5.2 節でドキュメントから重み付けを行って抽出したキーワードとの関連性を「4.3 タグ・キーワード間の関連性評価」で述べた手法に基づいて計算した結果を降順に並べた上位5つを表したものが以下の表 3 である．

表 3 付与された各タグのコンテンツ内容との関連度計算結果

1	**プログラミング言語**	**2.666**
2	javascript	2.611
3	**java**	**2.527**
4	ウェブ	2.404
5	**オブジェクト指向**	**2.280**

5.2 節と同様に評価方法は，再現率と適合率を用いる．候補集合を「プログラムが対象ドキュメントのコンテンツ内容を表すタグであると予想した集合」，正解集合を「人間がコンテンツ内容を表すタグであると判断した集合（表 3 において太字で示した語彙）」とした時の再現率と適合率のグラフを以下の図 4 に示す．

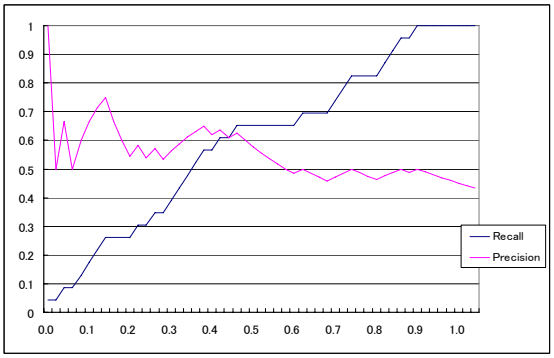

図 4 コンテンツ内容を表すタグの抽出結果

[*1]「初心者がJavaを超高速で学ぶためのコツ」
http://itpro.nikkeibp.co.jp/article/COLUMN/20060927/249181/

5.4 各語に関する関連概念抽出実験と結果

5.2 節の重み付けにより得られた結果の上位 93 語のキーワードと，4.1 節において人手により「①対象 URL のコンテンツ内容のキーワード・主旨を表すもの」と分類された中の Wikipedia クレンジング後に残った 22 語のタグを対象とし，重複する 5 語を排除した全 110語について関連概念を求める．

以下の表 4 が 4.4 節で示した手法を用いて得られた結果の1つ「プログラミング」についての関連概念を抽出結果を降順にソートしたものの上位 10 位である．得られた関連語彙が対象文書中に含まれていたものを「Keyword」，対象文書へ付加されたタグを「Tag」とする．また，同じ対象ドキュメントに対して従来のドキュメント中に登場する語の共起関係から関連語彙を求める手法である WordSpace により得られた結果も示す．

表 4　「プログラミング」に関する関連概念抽出結果

	Wikipedia	Keyword/Tag	WordSpace
1	プログラミング言語	Tag	COBOL
2	コンピュータ	Tag	構造化プログラミング
3	プログラマ	Keyword	書
4	構造化プログラミング	Keyword	Java
5	アルゴリズム	Tag	経験
6	オブジェクト指向	Keyword , Tag	勉強
7	ソフトウェア	Tag	学習
8	ソースコード	Keyword	理解
9	配列	Keyword	例外処理
10	変数	Keyword	例外

5.5 考察

5.2 節，5.3 節いずれも閾値の設定によっては，precision と recall の値が高く F-measure の値も 65．0%以上の結果が得られ良い結果が得られた．したがって，対象文書のコンテンツ内容を表すキーワードに関して Wikipedia マイニングの語彙間の関連度結果を用いることで抽出が可能であり，コンテンツ内容を表すタグに関しても「①対象 URL のコンテンツ内容のキーワード・主旨を表すもの」を高い精度を持って抽出可能であることがわかる．

5.4 節において，表 4 の「プログラミング」に関する上位関連概念の結果を見てみると，Wikipedia により得られた上位の関連概念は，例えば1位の「プログラミング言語」は「プログラミング」と has-a 関係を持つといった具合に，プロパティ定義が比較的容易にできる．また，タグから得られた関連語彙の高い確率で上位に登場していることがわかる．一方，語の共起性による評価（WordSpace）では，「書」，「経験」など，概念定義をする際ゴミとして扱ってよい概念が上位に登場してしまっている．関連性の低い概念が上位に登場してしまうことは概念定義のコスト増大を意味する．このような結果が得られた原因として，対象ドキュメントのコンテンツ自体が少なかったことがあげられるが，同時にこれは Wikipedia をリソースとして関連度を計った場合，語の共起関係を正しく得るために必要な膨大な量の専門文書等を対象としなくても概念定義が可能であることを示している．より詳細に従来の手法との差を検討するために，タグから得られた 17 語を排除し，テキストから得られた 93 語のみに対して候補集合を「プログラムが概念定義可能であると予想した集合」，正解集合を「人間が概念定義可能であると判断した集合」とした時の再現率と適合率のグラフを図 5 に示す．本研究で用いた語彙間の関連度は日本語 Wikipedia 内で完全にリンク関係の無いものであったら関連性が 0 となるため「プログラミング」以外の 92 語中 42 語までしか関連度が計算できない．しかし，図 5 の本研究の手法（実線）と従来の語の共起性による手法（破線）の再現率と適合率を見てもわかるように，Wikipedia によって関連性の取れた 42 語に関してはその中の 35 語の 83.3%が正解であり，従来の手法と比べ非常に精度の良い結果が得られた．以上より，元々Wikipedia という汎用的であるリソースを用いて，コンテンツ内容のキーワードと付与されたタグを評価することで，領域オントロジーを構築する際の概念定義支援が可能であることが示された．

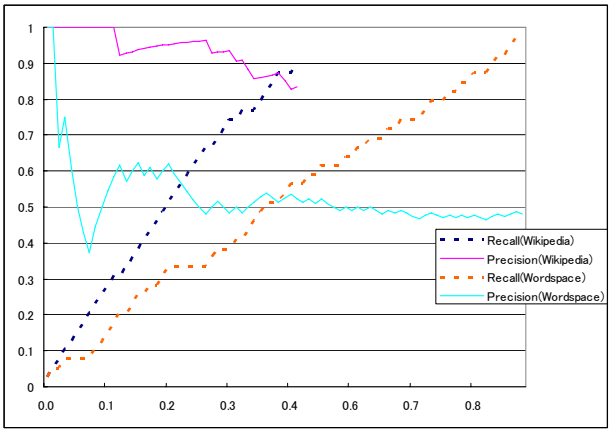

図 5 関連概念抽出結果の従来の精度との比較

6. おわりに

本研究では，汎用的かつ半構造情報資源である Wikipedia と Folksonomy タグに基づく領域オントロジー構築支援の提案を行った．ある IT 関連の文書を対象にしたケーススタディを通して，Wikipedia と Folksonomy タグに用いられる語彙に Wikipedia マイニングに基づいて得られた語彙間の関連度を利用した概念定義の有用性を示した．現在も日本語 Wikipedia は急速なスピードで成長をし続け，2007 年1月現在の総記事数は約 31 万記事と膨大であり，さらに今後，各領域に対する専門用語の網羅率は高まることが予測される．しかし，Folksonomy によるタグ付けの情報に関して今回利用したものはほんのわずか付加的なものである．フィード等を利用し，現在より有益な情報を活用していく方法を模索中である．今後の展望としては，より多くの特定領域に関するドキュメントを対象にオントロジー構築を行い，混沌としたタグ等の Wikipedia でさえも網羅されていない語彙の扱い，また，概念間の階層関係定義について検討していく必要がある．

参考文献

[Morita 06]T.Morita, N.Fukuta, N.Izumi and T.Yamaguchi: DODDLE-OWL: A Domain Ontology Construction Tool with OWL, First Asian Semantic Web Conference, LNCS4185, pp.537-551 (2006.9)

[Getoor 05]Getoor, L.and Diehl, C.P.: Link mining:a survey,SIGKDD Explorations,Vol.7,NO.2,pp.3-12(2005)

[中山 06]中山浩太郎，原隆浩，西尾章治郎著: Wikipedia マイニングによるシソーラス辞書構築手法，情報処理学会論文誌，Vol.47, No.10, pp.2917-2928,(2006)

[chasen]http://chasen.naist.jp/hiki/ChaSen/?FrontPage