1A3-04

World Wide Webを用いた辞典システムの構築

Construction of a dictionary system using World Wide Web

土田 正明[*1] 松井 藤五郎[*2] 大和田 勇人[*2]
Masaaki Tsuchida Tohgoroh Matsui Hayato Ohwada

[*1]東京理科大学大学院 理工学研究科 経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

[*2]東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

In recent years, people can access much information by spreading the Internet. However, it is inconvenient for a user with little experience to investigate about an unknown term. In this paper, we describe how to construct a dictionary system which extracts a term explanation note from a World Wide Web. We designed this system assuming that (1) the explanations of a term has certain fixed from and (2) similar words are used in explanations. We show the experimental results using words in computer science.

1. はじめに

従来，未知の用語を調べることには紙媒体の事典が用いられて来た．しかしながら，現代の情報社会においては，新しい用語が日々生み出されているにもかかわらず，紙媒体の事典では更新が困難であり役に立たないのが現状である．

一方，近年のインターネットの普及と情報検索技術の発展により，誰もが情報に用意にアクセスできるようになり，World Wide Web (WWW) から新語や専門用語に関する情報を比較的用意に取得できるようになった．

しかしながら，WWW から検索された情報には，検索された情報に目的の語とは関係ないものが数多く含まれるという問題点が存在する．したがって，それらの情報をさらに洗練する必要がある．

そこで本研究では，WWW の更新速度と情報の網羅性に着目し，目的の語を入力するとその説明や定義を WWW から抽出する辞典システムを構築する．筆者らはこれまでに，単一のページに説明文が含まれている物を対象にし，研究を行ってきた [2].

しかしながら，説明文は複数のページに存在している．そこで，本論文では単一のページからの抽出結果にこだわらず，情報を整理し，提示するシステムの構築を目指す．本システムの特徴は， 複数ページから抽出した文から，類似文判断による説明文の特定を行うことである．

2. 辞典システムの概要

本節では，本研究で構築する辞典システムの概要を示す．まず，本研究では説明文を判別するために，1) 用語の説明は一定の形をしていて，2) 説明文に使われる単語は類似しているという二つの仮定を行った．

本システムの構成図を図 1 に示す．本システムは検索エンジン，抽出モジュール，優先度付与モジュールの 3 つのモジュールから構成される．

本システムは入力用語に対して，検索エンジンにより文書を取得する．次に，それぞれの文書から用語説明文テンプレートによる抽出モジュールを用い，一次候補を生成する．最後に，一次候補の中からそれぞれの説明文を用い，優先度付与モジュールを用い，出力を決定する．以下にそれぞれのモジュールの説明を記す．

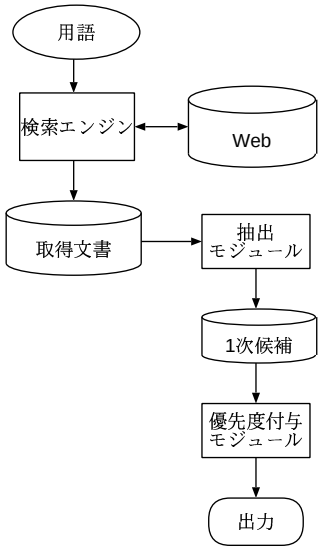

図 1: システム構成図

2.1 検索エンジン

本システムでは検索エンジンとして，Goo[*1]を用いる．

本システムに入力された語を A とすると，Goo に対し「A」,「A とは」,「A は」という検索要求を行い，それぞれ上位 100 件の検索結果のアドレスを取得する．100 件存在しない場合は全てを取得する．ここで重複は取り除くため，実際には 300 件にならないことに注意する．

最後に取得したアドレスから，pdf, ps, doc, ppt, swf などを除き，HTML や plain text の文書を取得する．

連絡先: 土田正明，東京理科大学大学院 理工学研究科 経営工学専攻，278-8510 千葉県野田市山崎 2641，04(7124)1501，masaaki@ia.noda.tus.ac.jp

∗1 http://www.goo.co.jp

2.2 抽出モジュール

検索エンジンによって取得した文書から HTML タグを取り除き，テキストのみを抽出する．そこで，本システムで行った仮定 1 にしたがって，用語説明文テンプレートを用い抽出を行う．

これまで類似研究として用語説明文のテンプレートを用いた物に，藤井らの研究 [4] と桜井らの研究 [1] がある．

藤井らの研究で用いられたテンプレートは，電子辞書から半自動的に生成した物であり，「A とは B」「A は B です」といった簡潔でわかりやすいものである．ここでは，それぞれのテンプレートは同列で扱われ，当てはまれば説明文であるとことより，意味的な違いは考慮されていない．

一方，桜井らの研究では，長尾らの研究 [3] を元に，用語の定義を表す情報の種類 (直接的内包，間接的内包，略記，目的など) ごとに，テンプレートを生成している．それらのテンプレートをより多く適用できるテキストの断片をシステム出力としている．

本研究では，複数のページに点在する用語に関する情報を統合し，出力することが目的となっているので，桜井らの研究で用いられたテンプレートを使用し，１次候補として出力する．

具体的には，検索用語が含まれる文から足していった結果，m 文字のはじめて超える文までをつの塊として扱い，それぞれの文にテンプレートを適用する．これは，あまりにも検索用語と離れた場所にある文は，その用語の情報ではない可能性が高いと言う仮定に基づいている．現在は経験的に $m = 400$ としている．

本システムでは，それぞれの定義の種類により，別々に一次候補として保持し，優先度付与モジュールへの入力とする．

2.3 優先度付与モジュール

本モジュールは,「説明文に用いられる単語は類似している」という本システムの仮定にしたがって構成される．本モジュールは一次候補の中で説明の情報を含んでいる物を選出する機能を果たす．

本機能を果たすために，本研究ではコサイン類似度を用いている．類似度を求めるために，本システムの開発言語である Java により実装された形態素解析エンジン Sen[*2]を用い，名詞，動詞，未知語を取得し，その全てを用いベクトル空間を生成する．

抽出モジュールにより作られたそれぞれの種類における全一次候補の対とのコサイン類似度を算出する．求めた自分以外との対の類似度の総和を優先度として使用する．

コサイン類似度は，同じ単語が同じ回数使われればそれだけスコアが高くなる (最大は 1.0) という特徴がある．また，複数のページに点在する情報に使われる単語が類似しているほど情報の信頼性も高くなるというメリットがある．

コサイン類似度を用いる事により本モジュールは，それぞれの情報の種類においてよく使われる単語が多く含まれるほどスコアが高くなり，その種類の代表として選出すると事が出来るという事がわかる．

現在，本システムの出力は，それぞれの情報の種類の上位 3 件としている．

3. 評価実験

本論文で提案した辞典システムの性能を評価するために．それぞれの情報の種類による精度を計測した．

*2 http://yamaguch.sytes.net/~tora/opensource/sen/

表 1: テンプレート

直接的内包	.*と?は.*(だ｜である｜です｜であり)*.*
間接的内包	.*と (い｜言)+(い｜う｜われる)+.*
略記	.*の (略｜略称｜頭文字｜略語｜訳語)+.*

表 2: 実験結果

	精度
直接的内包	47/76(0.605)
間接的内包	17/26(0.654)
略記	16/43(0.372)
全体正解	25/30(0.833)

本実験における評価のためのテストデータには，基本情報処理技術者のテキストにおいて用語として解説されていた語の中から適当に取り出した 30 語を用いた．

今回は情報処理技術の用語を対象にしたため，その用語問題に答える為に主に重要であると考えられる分類のうち,「直接的内包」「間接的内包」「略記」を対象として行う事とした．それぞれが用語を説明する情報であると判断される場合を正解とした．

それぞれの種類について桜井らの定義 [1] にしたがって簡単に説明する．直接的内包とは，非定義語がトピックとなる物であり，間接系内包とは非定義語が連体修飾句を構成である．また，略語は略称を示すものである．

本実験に用いたテンプレートを表 1 に示す．は任意の一文字，*は 0 回以上の繰り返し,+は一回以上の繰り返し，?は 0 回または一回の繰り返し，(A｜B) は A か B にマッチすると解釈する．

実際に本システムによって出力された物を例にとり具体的に説明する．本システムに「XML」を入力した場合の直接的内包による出力を以下に示す．

直接的内包の出力

1. XML.ORG は、OASIS サイト (www.oasis-open.org) とは別に XML.ORG サイト (www.xml.org) を持ち、XML ポータルをとしての役割を担っている．
2. XML では、HTML4.0+CSS と同様、XML+XSL として、文書の内容と表現 (書式) は分離して管理できます．
3. XML のメリットと利用形態： XML は，1986 年 ISO で標化された SGML(Standard Generalized Markup Language) をインターネットで活用しやすくするために，1998 年 2 月にその基本仕様 XML1.0 が W3C(World Wide Web Consortium) にて策定された．

本実験では 1 は XML 自体の特徴や性質を記述していないため不正解とし，2 と 3 は特徴や性質が記述されているため正解とする．1 の後半部「XML ポータルをとしての役割を担って

いる.」は原文が間違っていることに注意する.

間接的内包の例は以下の通りとなる.

間接的内包の出力

1. XML を使えば何ができるのか、という観点から参考になると思います.
2. XML に「拡張可能な（Extensible）」という名前が付けられた理由はここにある.
3. XML が非常に重要な技術である理由、それは、ソフトウェアとデータの関係を変える技術であるということです.

直接的内包と同様の基準で 1 は不正解，2 と 3 は正解となる．略記に関しては「XML とは、eXtensible Markup Language の略で、直訳すると拡張可能なマークアップ言語となります.」といった略語が一目瞭然な物のみを正解とした．

また，全てを読めば，用語が何であるかがわかると判断した場合の精度も調べた．これを本論文では「全体正解」と呼ぶことにする．その場合の正解の基準は，主観的な判断により，全てを読めば，基本情報処理技術者の用語問題に誰が読んでも正解できると判断された場合に正解とした．上記の XML を例にとれば，文から「拡張可能な」「eXtensible Markup language」「SGML の後継」「W3C により策定」などから十分理解可能であると判断され正解となる．

結果を表 2 に示す．略記に関しては略記が存在しない用語がある事や略記を記述していない文もテンプレートに多く当てはまっていたため，多少精度が低い結果となったが，直接的内包に関しては，用語あたり，平均 1.567 個の正解を含んでおり，結果的に全体正解の 0.833 という高精度に結び付いたと考えられる．

4. 考察

実験の結果からわかるとおり，全体正解の精度は 0.833 と高い精度となっているが，直接的内包，間接的内包，略記などはどれもよい精度とは言えない．

これはテンプレートを変更する方法と，優先度付与モジュールを改善する方法があると考えられる．その検証のため，全体正解をしなかった 5 つの用語について直接的内包，間接的内包，略記をそれぞれ調べた．

その結果，5 つの用語の内 2 つの用語は正解が含まれているにもかかわらず優先度付与モジュールによって弾かれていた事がわかった．正解が含まれていない場合の傾向として，テンプレートによりマッチした文の数が少ない事が特徴として現れた．よって，候補が多い場合は優先度付与モジュールが情報を絞るために有効に働いていると言える．

3 つの用語に関しては，本システムでは一次候補の段階で正解を抽出していない結果となった．これはテンプレートそのものの問題もあるが，検索用語に対する用語解説ページには見出しになっているだけで，本文中には検索用語が含まれていないケースもあることがわかった．この点に関しては，HTML タグの構成を見ることなどによって改善の余地があると考えられる．

以上の結果から，テンプレートを少し緩め，一次候補を増やすか，文中に検索用語がない場合に対応するなど，網羅性を高めれば本システムの優先度付与モジュールが有効に働くと考えられる．

5. まとめ

本稿では，WWW から用語の説明や定義を抽出する事典システムを構築し，その評価実験を行った．

本システムは抽出モジュールと優先度付与モジュールから成り，抽出モジュールは検索結果からテンプレートを用いて説明文の候補を抽出し，優先度付与モジュールが候補に優先度をつけて説明文を選択する．

実験の結果，その用語の上位概念と，いくつかの特徴を含んでいるという評価基準において 0.833 という高い精度を実現した．本システムを用いれば，基本情報処理技術者の用語問題に答える事が出来るという実用性も確認できた．

本システムはどのような分野の用語にも対応しているので，説明書の用語説明の自動生成やオンライン教材の自動生成などが幅広い応用が考えられる．

参考文献

[1] 桜井 裕， 佐藤 理史: ワールドワイドウェブを利用した用語説明の自動生成, 情報処理学会論文誌, Vol. 43, No. 5, pp. 1470–1480, 2002

[2] 土田 正明， 松井 藤五郎, 大和田 勇人: World Wide Web からの検索を用いた辞典システム構築のための文書分類, 情報処理学会第 66 回全国大会講演論文集, Vol. 3, pp. 83–34, 2004

[3] 長尾 真: 辞典形式での専門分野の知識の体系的構成方法, 人工知能学会会誌, Vol. 7, No. 2, pp. 336–345, 1992

[4] 藤井 敦, 石川 徹也: World Wide Web を用いた辞典知識情報の抽出と組織化, 電子情報通信学会論文誌, Vol. J85-D-II, No. 2, pp. 300–307, 2002.