英文テキストからの索引自動生成に関する研究

6S－2

植松純一　原田賢一
慶應義塾大学理工学部

1　はじめに

部品の検索の方法は、適用されているアプローチから大まかには、知識ベースを用いた人工知能による方法 [1, 3] と図書館情報学から発達した情報検索による方法 [4, 5, 6] に分類できる。私は、情報検索の方法の方が、有力であると考える。

情報検索による方法には, 整理されていない語を用いる方法 (Uncontrolled Vocaburary) , 項目と値を用いる方法 (faceted) , 列挙する方法 (enumerated) , キーワードを用いる方法 (keyword) , 属性と値を用いる方法 (attribute-value) などがある。[5] の研究によれば、前述のように様々な方法が存在するが、方法の違いによる正確さ、召喚率の違いは大きくない。また、人間の好みについても、大差ない。したがって、部品について必要な情報をいかに抽出するかが、問題である。

本研究では、部品の情報を自動的に抽出することすなわち、自動的に索引を生成することを考える。これは、ライブラリの保守を容易にし、さらに、適切な抽出を行うことは、正確さと再現率を高めることになるので、大変意義あることといえる。

2　素材

部品の索引を自動的に生成する素材となる候補としては、ソースコード自体、ソースコードに挟まれたコメント、付属するマニュアルなどのドキュメント、設計時の設計仕様書などが考えられるが、マニュアルなどのドキュメントを、索引を生成する素材として用いることにする。

素材として、マニュアルなどのドキュメントすなわち、自然言語を対象としているので、電子化された様々な文書などに応用出来ると考える。

3　情報自動抽出に関する今までの研究

項目と値を用いる方法 (faceted) や属性と値を用いる方法 (attribute-value) などの検索を実現しようとする際に、部品に関する必要な情報を、自動的に作成することは大変困難である。列挙する方法 (enumerated) を実現するには、階層構造を自動的に作成することになる。これに関する研究もあるが、適切な階層構造はなかなか得られていない。

古くからある部品の情報を自動的に抽出する方法としては、キーワードによる検索を前提として、部品に付属するドキュメントなどから、その部品に関するキーワードを抽出することがなされてきた。しかし、十分な成果はあげていない。[2]。

部品の情報を一単語としてではなく、二単語として取り出すことを考えた研究もある。取り出した二単語をLA(Lexical Affinities) と呼ぶ [6]。ある一つの部品に対して、複数のLA で構成されたプロフィールを作成し、これをもとに検索を行う方法が研究されている。

4　本研究のアプローチ

4.1　適用する検索方法

キーワードによる方法を考える。キーワードによる方法を考えた場合、一語による検索では、能力が低い。そこで、[6] で研究されたLA をキーワードと捉え、検索することを考える。二単語で構成されるLA をキーワードとして扱うことにより、部品の表現能力が向上する。

4.2　LA の雑音除去、重み付け

従来の研究では、LA は文書中の頻度により抽出され、基本的に頻度の多いもの程、重く重み付けされていた。

しかし、頻度の多いものほど、その部品を表すプロフィールとして適切であるとは限らない。

従来のLA では、単語間の関係について解析を行うことはなく、ある単語から5 以内の単語を全てLA と考えて、LA を抽出してきた。しかし、これでは、LA として取り出すべきでない全く関係のない2 単語をLA として取り出すことになる。

そこで、ドキュメントなどの素材の単語一つ一つについて品詞を解析する。そして (動詞、名詞)、(名詞、名詞) (形容詞 、名詞) の組合せのものについてだけ、LA として認識して取り出す。この作業により、かなりの雑音が除去できる。ドメインの情報を使うより正確な雑音除去が出来る。

Automatic Generation of Index from English Text
Junichi UEMATSU, Ken'ichi HARADA
Keio University
3-14-1 Hiyoshi, Kouhoku-ku, Yokohama 223, JAPAN

これらを考慮にいれて、LAの集合体であるプロフィールを作成する研究を行なう。プロフィールは、複数のLAで構成されるわけであるが、それ自体を部品を表現するものとして探索時に使用するので、LAの重み付けが重要になる。

品詞が特定されているので、名詞の複数形や動詞の活用は全て原形に出来る。

LAの重み付けをする際には基本的には頻度により重み付けするが、何をどうするという情報を運んでいる(動詞、名詞)の組合わせならば、二倍の重みがあると考えて処理する。

検索時には、一つのLAに対して複数の合致する検索対象が存在する。これを重み付け大きさの順に表示することは適当ではない。検索対象のなかで、求めるLAが何番目に重要なLAとして重みずけされているかを調べて、それに準じて表示する。

5 本研究の実装例

本研究の実装例として、UNIXのコマンドを検索対象と捉え、そのMANページを情報を抽出する素材として、プロフィールを作成する実験を行った。この実験のために作られたプログラムは、英文のテキスとなら何でも適用できるプログラムである。また、ドメインに依存する特別な処理は行っていない。

従来の方法と比較するために、4つのプロフィールを作成した。

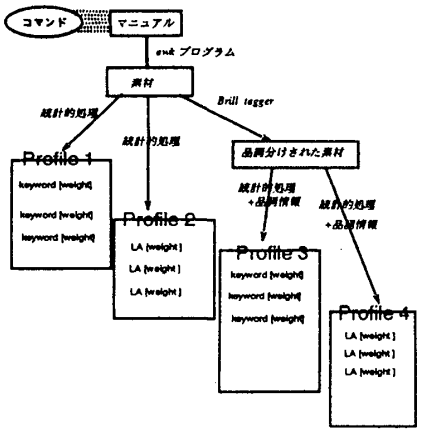

図 1: 実装の実際

インターフエースもWWW上に実現した。'http://www.hara.cs.keio.ac.jp/ contrib/junichi/shuron/dai.html'

6 今後の課題

LAの重み付け、雑音除去を最適なものにするために、さらなる分析が必要である。

キーワード探索の問題として、ある一つのものを表すのに人によって違った単語を使用する問題がある。そのために、類義語辞典を使用して、この問題を解決する方法を考える。

検索技術に求められる性質として、似た部品を検索する必要がある。よって、拾い読み(Browsering)を支援する仕組みを提供出来なくてはならない。そのためには、部品をグループにまとめる必要があり、部品の類似性を測定するアルゴリズムを研究しなくてはならない。

他のドメインに対して応用してみる必要がある。

7 まとめ

キーワード的検索を実現するために、LAを改良したものをキーワードとして用い、品詞を意識して雑音除去、重み付けを行なった。また本研究の適用例として、UNIXのコマンドを検索対象としてとらえ、検索データベースを実現した。そして、今後の課題を解決することが必要である。

参考文献

[1] Eduardo Ostertag , James Hendler ,Ruben Prieto Diaz and Christine Braun : Computing Similarity in a Reuse Library System :An AI-Based Approach, ACM Transaction on Software Engineering and Methodology, Vol.1, No.3 (Jul. 1992), 205–228.

[2] G.W.Furnas,T.K.Landauer,L.M.Gomez and S.T.Duma : The Vocabulary Problem in Human-System Communication, Communication of the ACM, Vol.30, No.11 (Nov. 1987), 964–971.

[3] Premkumar Devanbu,Ronald J.Brachman,Peter G.Selfidge, and Bruce W.Ballard: LASSIE:A Knowledge-Based Software Information System, Communication of the ACM, Vol.34, No.5 (May, 1991), 35–49.

[4] Ruben Prieto-Diaz and Peter Freeman :Classifying Software for

Reusability IEEE Software,(Jan.1987),6–16.

[5] William B.Frakes and Thomas P.Pole: An Empirical study of Representation Methods for Reusable Software Component, IEEE Transaction on Software Engineering, Vol.20, No.8 (Aug. 1994), 617–630.

[6] Yolle S.Maarek,Daniel M.Berry,and Gail E.kaiser: An information Retrieval Approach For Automatically Constructing Software Libraries, IEEE Transaction on Software Engineering, Vol.17, No.8 (Aug. 1991), 617–630.