1G2-02

情報の深さを考慮した情報獲得支援システム

Information Acquiring Support System Using the Depth of Information

西原 陽子 *1　　砂山 渡 *2　　谷内田 正彦 *1
Yoko Nishihara　　Wataru Sunayama　　Masahiko Yachida

*1大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University

*2広島市立大学情報科学部
Faculty of Information Sciences, Hiroshima City University

With development of WWW, we more often acquire information using search engines. Although there are many useful information that are contained in web pages, some of them are too difficult to understand by the reason the used technical terms are too much special. So, to get understandable information, we have to search again out of the search results. We propose a system which presents the web pages according to a user's knowledge level. In this paper, we consider that a web page is a document written about one theme. This system gives the degree of speciality to all the technical terms in documents and gives the depth of information that means the information's difficulty to all the documents using degrees of speciality. From experimental results, it proved that this sytem can give suitable difficulties to web pages.

1. はじめに

WWW の普及により，検索エンジンを用いて情報獲得をする機会が増加した．WWW 上には膨大な量の web ページが存在し，その中には有益な情報が多数含まれているが，使われている専門用語がユーザの知識レベルを超えているために内容が難しくて理解できないものも存在する．そのため，ユーザは得られた検索結果の中からユーザが理解できる web ページを再検索する必要がある．

情報獲得支援を目的として，AreaView2001[1] はある分野に関連した web ページを整理・組織化することによってその分野の外観を把握する「入門書」を作成してユーザの情報獲得を支援しているが，ユーザの知識レベルを考慮しておらず再検索の必要性の問題は解消されていない．

そこで本稿では web ページを１つのテーマについて書かれたドキュメントと見なし，その中で用いられている専門用語の専門度から各ドキュメントにその情報の深さを示す難易度をつけ，ユーザの知識レベルに適合したドキュメントを提示する手法を提案する．ドキュメントはその内容でクラスタリングを行い，ドキュメント間の関連を明確にしてユーザに提示する．本手法によって効率良い情報獲得の支援を図る．

2. 提案システムとドキュメントデータベース

提案システム構成を図 1 に示す．本システムは知りたい情報を示すキーワードとユーザの知識レベルを示す既知の専門用語を入力とする．システムはドキュメントデータベース（以下データベース）から，全クラスタの関連を示したクラスタマップと入力キーワードに関連があるクラスタ内のドキュメントをユーザの知識レベルに近い順に並べて出力する．本システムでは１つのクラスタ内のドキュメント内容を習得したら，クラスタマップを参照して現在のクラスタに隣接する新たなクラスタへ移動し，知識を深めつつ情報獲得を行うことができる．

ここで本システムで獲得できる情報はデータベース内に存在する分野に限定されている．データベースにはあらかじめクラスタリングを行い，難易度がついた状態でドキュメントが

連絡先: 西原陽子，大阪大学大学院基礎工学研究科，560-8531 豊中市待兼山町 1-3，tel: 06-6850-6363，fax: 06-6850-6341，yoko@yachi-lab.sys.es.osaka-u.ac.jp

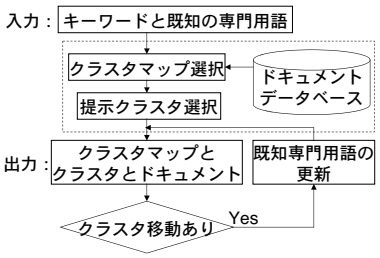

図 1: システム構成

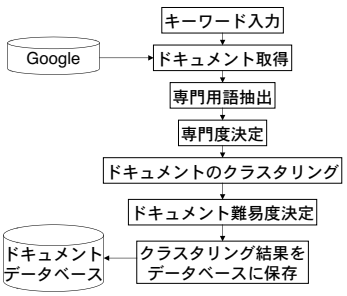

図 2: データベース作成手順

保存されている．データベース作成手順を図 2 に示す．まず，知りたい情報の分野を示すキーワードを Google[2] に入力し，検索の結果得られるドキュメント（web ページ）を取得する．続いてドキュメントから専門用語を抽出し専門度をつける．次に全専門用語をクラスタラベルとして，ドキュメント中にクラスタラベルを含むものをまとめてクラスタリングを行い，ドキュメントに専門度から難易度をつけ，難易度の低い順に並べ替えてデータベースに保存する．クラスタリングは内容の似ているドキュメントをまとめることで，ドキュメント間の関連を分かりやすくするという効果がある．

2.1 ドキュメントの定義

本システムにおいてドキュメントとは１つのテーマについての説明が記述された説明文と定義する．ドキュメント中には

テーマを表す専門用語とテーマを説明ための専門用語が含まれている．

2.2 専門用語とその専門度の定義

専門用語はドキュメントのテーマを表すキーワードとテーマを説明するために用いられているキーワードと定義する．キーワード抽出には展望台システム [3] を用いた．展望台システムは文章に観点語が与えられると，共起確率から新たな観点語を補完し，観点に合った文章を特徴づける特徴語と文章中での頻度が高い語を背景語として抽出し，観点語，特徴語，背景語を用いて文章から重要文を抽出するシステムである．文章を 1 つのテーマについての説明文，テーマを表すキーワードを観点語とした場合，抽出される特徴語はテーマの説明に使われているキーワードであると考えることができるので展望台システムによる専門用語抽出を行った．

専門用語は専門度の高いものほど理解しにくい．本システムにおいて専門度は専門用語を理解するために必要な事前知識量と定義する．必要な事前知識量が少ないならば専門性が低いのでドキュメントテーマの説明にも多用される可能性が高く，含まれるドキュメント数も多くなると考えられる．したがって，all を対象としている全ドキュメントの数とすると，専門用語 $term$ が含まれるドキュメント数 $n(term)$ から専門度 $s(term)$ を測ることができる．

$$s(term) = \frac{all - n(term)}{all} \tag{1}$$

また，異なる専門用語 $t1$ と $t2$ があり，それぞれの専門度が $s(t1) > s(t2)$ で与えられる時に $t2$ が $t1$ に依存しているならば $t1$ の理解に必要な事前知識量で $t2$ の理解に必要な事前知識量の一部を負担できると考えられる．すなわち，両専門用語を同時に理解するために必要な事前知識量はそれぞれの専門度の和よりも少なくなる．ここで，$t2$ が $t1$ に依存していない確率は

$$P(\bar{t1}|t2) = \frac{P(\bar{t1} \cap t2)}{P(t2)} = \frac{n(\bar{t1} \cap t2)}{n(t2)} \tag{2}$$

と表せ，$t2$，$t1$ の同時専門度 $s(t1, t2)$ を

$$s(t1, t2) = s(t1) + s(t2)P(\bar{t1}|t2) \tag{3}$$

で表す．

2.3 ドキュメント難易度の定義

ドキュメント難易度はそのドキュメントを理解するために必要な事前知識量と定義する．ドキュメント内容を理解するためには、含まれている専門用語を理解している必要がある．ここで，難易度が高いドキュメントほどドキュメントテーマを示す専門用語の専門度は高く，テーマを説明するのに用いられている専門用語数は多く，専門度の高いものも多数含まれると考えられる．このときドキュメント中で最大専門度値をとる専門用語 t_{max} に着目するとドキュメント中の t_{max} 以外の専門用語のいくつかは t_{max} の説明に使われている可能性が高く，そのような語は t_{max} に依存している分専門度が小さくなると考えられる．したがって，ドキュメント D の難易度 $d(D)$ は

$$d(D) = s(t_{max}) + \sum_{t \in S} s(t)P(\overline{t_{max}}|t) \tag{4}$$

で測ることができる．S は D 中の専門用語から t_{max} を除いたものである．

正解難易度	ドキュメントの見出し語	式 (4) での難易度
易しい	RFC	7916(5)
	ウェルノウンポート	5671(7)
	FTP	2316(9)
	TCP	4067(8)
	TCP/IP	6378(6)
	プロトコル	14510(1)
難しい	マルチレイヤスイッチ	12487(2)
	レイヤ 7 スイッチ	11125(4)
	ファイル共有	12286(3)

表 1: 正解難易度毎の FTP クラスタ内ドキュメントの難易度（カッコ内はその順位）

3. 評価実験および考察

式 (4) によってドキュメントに適切な難易度が付与できるかを確認する実験を行った．ドキュメントは IT 用語辞典 [4] のネットワーク技術カテゴリに属するネットワークに関する用語の説明文を 406 個用意した．説明文の見出し語をドキュメントテーマを表す専門用語として，２章で説明した手順に従いデータベースを作成した．クラスタラベルは 1414 個でその中から「DNS」,「FTP」,「MAC」クラスタ内のドキュメントに対して人手で正解専門用語を抽出し，正解専門用語からドキュメントに「易しい」,「難しい」の２値で正解難易度をつけ，式 (4) の難易度と比較した．

表 1. において「マルチレイヤスイッチ」と「ファイル共有」の難易度は前者の方が上である．後者の方が高い専門度値を持つ専門用語が多かったので，難易度を単純に専門度の和で表すと後者の難易度の方が高くなる．しかし、後者のドキュメント中に含まれる専門用語は他のドキュメント中において共起する確率が高く，それぞれ個別に理解する必要はないためにドキュメント内容は分かりやすいものであった．また，前者の方がより専門的な語でありそれを説明している文の真の難易度は高くなると考えられる．式 (4) によって関連度の強い専門用語の専門度を抑えて難易度を求めることができ，ドキュメント難易度をより精確につけることができた．

4. まとめ

本稿では，ドキュメントの専門用語の専門度からドキュメント難易度をつけ，ユーザの知識レベルに適合したドキュメントを提示する手法を提案した．評価実験から提案手法によってドキュメントに適切な難易度がつけられることを確認した．

今後の課題としては展望台システムによって抽出される特徴語がドキュメントの専門用語として妥当なものであるかの検証，クラスタリング方法の検討，ドキュメントの提示法の検討等がある．

参考文献

[1] 平博司，福島伸一，大澤幸生，伊庭斉志，石塚満：AreaView2001:WWW からの構造化した領域総覧提示システム，人工知能学会誌，Vol.17， No.3， pp.268-275(2002).

[2] http://www.google.co.jp/.

[3] 砂山渡，谷内田正彦：観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装，人工知能学会論文誌， Vol.17， No.1， pp.14-22(2002).

[4] http://www.itmedia.co.jp/dict/network/.