情 報 学 基 礎 44－3
（1996. 11. 21）

キーワードの位置情報を利用した情報検索技術の提案

羽田 久一　　山口 英

奈良先端科学技術大学院大学　情報科学研究科

内容梗概

従来より電子テキストの検索では、キーワードを与えそれにマッチするものを選びだすことが一般的に行われている。これらの方法ではキーワードのリストとそれに関係する論理式により、利用者の要求をシステムに伝えるが、論理式とキーワードだけでは単純なキーワードのマッチングを処理することしかできない。

本研究では従来のようにキーワードの存在により情報を検索するのみでなく、キーワードの位置情報をもとに利用者の要求する情報を取得する方法を考案した。また、この方法の利用分野を考察し、応用範囲についても述べる。

Information Retrieval Method using Keyword Location.

Hisakazu Hada　and　Suguru Yamaguchi

Graduate School of Information Science, Nara Institute of Science and Technology

Abstract

Traditional way of retrieving electronic text is using keyword matching search.

In this method, user inputs keywords and logical operators, which describes their relationship, to the system. However, this method only supports simple keyword matching.

This paper reports on the new search method using location of keywords and determination of document characteristics by using search result. It also discusses application of this method.

1 はじめに

ネットワークやマルチメディア技術の発達により、多種多様な情報の電子化および蓄積が行なわれている。そしてこれらの電子化された情報を一元的に提供する「ディジタル図書館」の実現に対する期待が高まりつつある。従来型の図書は電子化され、電子テキスト、あるいは画像情報として蓄積される。そのため紙媒体では考えられなかったような、柔軟な検索が可能になる。

現在一般的に行われている図書の検索方法は書誌情報などの二次情報を用いた検索であり、現存する図書館において、蔵書カード、あるいはMARC(Machine Readable Catalog) などの形式により検索を行うことができる。

電子図書館では、書誌情報など二次情報のみならず、電子化されたテキストを扱うため、一次情報を直接検索することができる。この電子化されたテキストを直接検索するものが全文検索と呼ばれる技術である。

全文検索では書誌のタイトルやキーワードのみならず、デジタル化された本文全体を検索対象とする。そのため本文中に存在する単語全てを検索対象とすることが可能であり、キーワードの選択ミスにより情報を逃すことが少なく、従来のMARC などでは検索できなかった書籍や文書を発見することができる。

しかしながら、全文検索では多量のテキストデータを検索対象とするため、検索速度の低下や無用な検索結果すなわちノイズの増大が避けられない。

そこで本研究では全文検索において、キーワードが文書中でどの位置に存在するかという位置情報を導入することにより、情報の検索におけるノイズの混入を低減し、的確な検索を行うことを提案する。また、数種類の電子化テキストに対してこの方法を適用し、効果を検証した。

2 位置情報を利用する検索

2.1 位置情報の重要性

文書中に単語を発見する場合、前後の語句やその単語の出現位置は非常に重要な情報を含んでいる。例えば論文の場合、実験や考察に含まれる単語は「はじめに」の部分のみに含まれる単語にくらべ重要度が高いと考えられる。

また一般的に文書は構造を持っており、その構造上重要な部分と重要でない部分に分けることが可能である。まとめなど重要な部分で多用されている語はその文書の内容を示す重要な単語であると認識することができる。

このように単語の位置情報に着目することによって、従来は自然言語処理の分野で考えられてきた、重要な単語の抽出や、文書の主題の抽出などを疑似的かつ、簡単な手法によって行うことが可能である。

同様のものとしては、UC berkeley における DL プロジェクトで、Tilebar と呼ばれるインターフェースが開発されている。[1] この Tilebar は文書の長さに応じた桝目 (タイルと呼ばれる) を用いて、検索結果を表示する仕組みである。キーワードを発見できた部分は桝目に色をつけて区別している。

2.2 頻度と位置情報

文書をページなどのある単位で分割する場合、文書の部分ごとの単語の出現数もその文書を理解する上での手がかりとなりうる。

ある文書を「インターネット」と「電子図書館」の2つのキーワードで検索する例を考える。従来の全文検索においては、これら2つの単語がマッチする文書を選択する。

ここで位置情報を利用するとインターネットに関する文書で電子図書館について述べられているものと電子図書館についての文書中でインターネットに関する記述が存在するものを明確に分類することが可能である。

位置情報を導入するために、この文書を同じサイズの部分 (ページ) に分割し、そのページごとに頻繁に出現する単語に関するヒストグラムを作成する。疑似的にヒストグラムを表した

ものを図1に示す。

ある単語が文書の広範囲にわたり出現している場合には、その単語は文書の主題の一つであると考えることができる。MAINTOPICは文書の全体に出現しており、この文書における主要な役割を果たすキーワードの一つであると考えられる。それに対してSUBTOPIC のように、ある単語が一部分に集中して登場する場合には、その単語は文書の副題であり文書中のテーマの一つであるが、文書全体に適用される主要なテーマではないと考えられる。

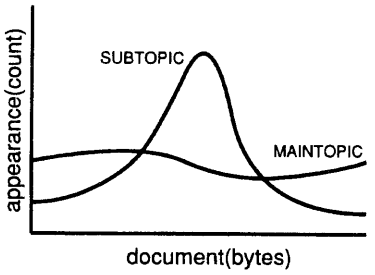

図 1: 単語の分布と主題、副題の相関

2.3 単語の出現パターンによる相違

「電子図書館」と「インターネット」の両方にマッチするような文書を選択した場合の、本手法における解析を行う。

位置情報を利用することにより、インターネットと電子図書館の双方が述べられている文書を選択するのみならず、「電子図書館に関する文書でインターネットについて述べたもの」と「インターネットに関する文書で電子図書館について触れたもの」を区別することが可能である。

図2ではインターネットを主題とする文書中で電子図書館について述べられているが、図3のようなグラフを得た場合には電子図書館が主題の文書中にインターネットに関する記述が存在したと考えられる。この2つの例はインターネット、あるいは電子図書館がメインのトピックとなっている文書である。

どちらも副題である場合には図4のように、ある文書中に別々にインターネットに関する記述と電子図書館に関する記述が存在する場合と、図5のようにインターネットと電子図書館の関連について述べる章を持つ場合とが存在する。

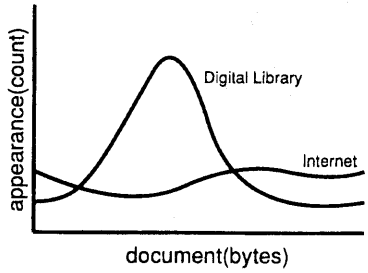

図 2: 主題と副題の分布 (その1)

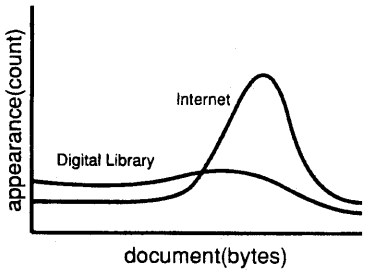

図 3: 主題と副題の分布 (その2)

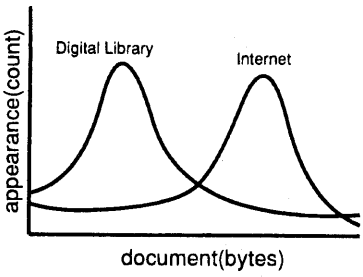

図 4: 関連のない副題

ヒストグラムを作成することにより、従来の「キーワード」と「論理式」で取り出した情報にも内容の違いがあることが判る。

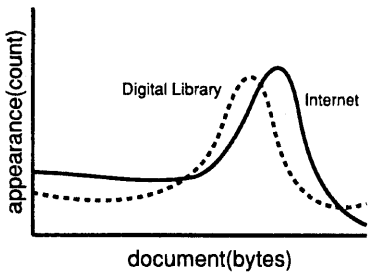

図 5: 関連する副題

2.4 DLA 法

単語の文書中での位置を記録し、そのデータを元に情報の検索や分類を行う方法を DLA (Data Location Analysis) 法と命名した。DLA 法では文書を単語に分割し、その単語が文書中のどの位置に属するかを記録する。この時、一般的に利用されキーワードにならないと考えられる語句は無視するよう辞書に登録されている。

文書から DLA 情報を作成し、DLA 情報から可視化を行う。可視化された状態の DLA 情報を DLA 地図と呼ぶ。利用者や管理者は DLA 地図を見ることにより、視覚的に文書の特徴を判断することができる。これが DLA 法の一番の特徴である。

3 実装と実験

3.1 実装

DLA 法の有効性を示すため、単語およびその位置の抽出と、ヒストグラム画像作成のプログラムを作成した。

プログラムは文書より単語の位置と頻度を抽出する解析器と、解析結果を利用し、DLA 地図を作成する可視化器にわかれている。さらに、HTML 文書を扱うために HTML タグを文書中から消去するフィルタを作成し、前処理として利用した。

これらのプログラムはすべて perl で記述されており、画像ファイルの生成には画像ファイル作成用パッケージである fly [2] を利用している。プログラム間はファイルを使い情報を伝達するため、新しい作業の追加や機能拡張を行いやすくなっている。

機種依存性を排してあるため、画像生成パッケージ fly および perl の動作する環境ならば、パッケージの動作が期待できる。今回の実験は SGI 社の計算サーバ Challenge XL 上にて行った。

3.2 実験

DLA 解析器および可視化器を用い、既存の文書に対して DLA 法を適用する実験を行った。

対象とする文書は WWW サイトより入手した Digital Library Initiative 関連の論文である。D-lib magazine と呼ばれるこの WWW サーバ [3] は 電子図書館研究のためのオンライン雑誌である。

本実験で用いたテキストは Dlib Magazine 7-8 月号に掲載された、DLI の経過報告に関する論文である。DLI はアメリカの NASA、NSF などが資金提供を行っているプロジェクトで全米より 6 つの大学が選択され、先進的な電子図書館へ向けての実験が行われている。

WWW 上で提供されるデータは内容に HTML タグを含むため前処理として別プログラムにより HTML タグを消去している。

HTML タグを除去された論文データは DLA 解析器にかけられ、ヒストグラムの元となる情報が抽出される。このとき、文書の分割数は 10 であり、元文書のサイズによらない。元文書のサイズを表 1 に示す。

文書名	サイズ (bytes)
CMU Infomedia	5932
Stanford Univ.	8968
UIUC DL	8119
UCSB Alexandria	16415
UC. Berkeley	16869
Univ. of Michigan	7139

表 1: 元文書のサイズ

解析器において文書は単語ごとに分割され、その出現位置と出現回数の総数がカウントされ

る。カウントされた単語は、その出現位置とともに、ファイルへ記録される。これがDLA情報となる。

DLA情報は可視化器によりDLA地図として画像化される。現在のDLA地図はいくつかのヒストグラムが重なったものとして表現されている。利用者や管理者は、このDLA地図を見ることによりキーワードの関連や文書の構成を直観的に知ることができる。

3.3 結果

D-lib Magazineにあるアメリカの DLI (Digital Library Initiative) の報告のページからDLA地図を作成したものを図6から図11に示す。DLA地図にはそれぞれのDLA情報のうち上位5つのものがグラフ化されている。このグラフを観察することにより、各プロジェクトの主眼としている目的が少しづつ違うのが直観的に理解できる。

4 DLA法の応用

4.1 地図の応用

DLA法ではDLA地図を見ることにより、文書中のキーワードの分布を知ることが可能である。キーワードの分布パターンより、主題、副題について推測することができる。この地図の近似により文書の自動分類を行うことが可能であり、今後さらに増大していくオンライン情報の自動分類において力を発揮すると考えている。またユーザに対しても地図を表示することにより、キーワードの分布を知る事が可能になり、情報検索を助けることができる。

4.2 DLA情報の利用

DLA地図の元となるDLA情報はキーワードと数値の列からなるデジタルデータであり、コンピュータ処理を容易に行える形式となっている。

従来では文書の意味を解析しグルーピングする作業は人間が文書を目で確かめ、人間の手で行うことがほとんどである。

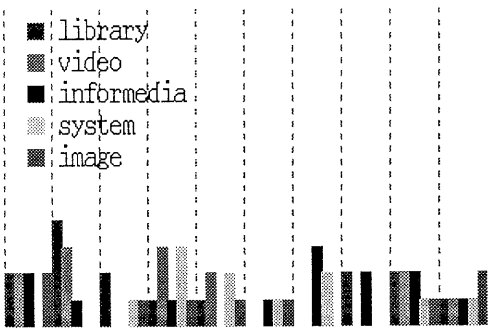

図 6: CMU Infomedia に関する論文の DLA 地図

図 7: Stanford Univ. DL に関する論文の DLA 地図

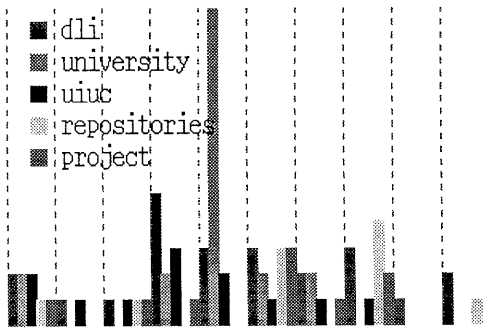

図 8: UIUC DL プロジェクトに関する論文の DLA 地図

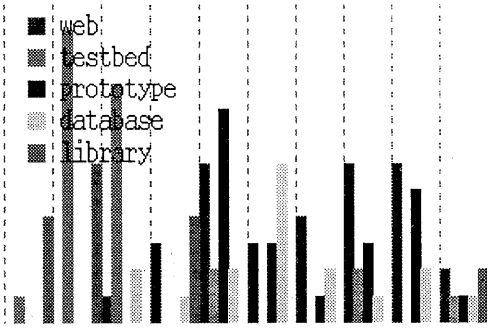

図 9: UCSB Alexandria 論文の DLA 地図

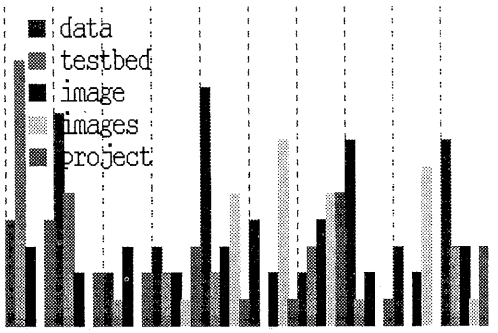

図 10: UC Berkeley DL プロジェクト論文の DLA 地図

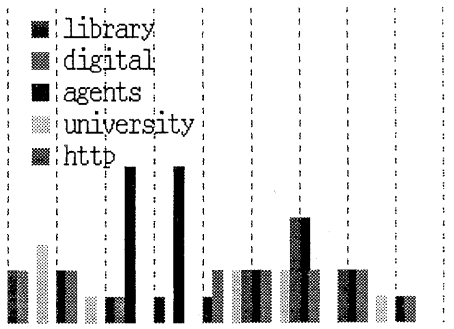

図 11: Univ. of Michigan UMDL 論文の DLA 地図

インターネット上ではほぼ無限といえるほどに情報が生産され、公開されている。yahoo[4]を代表とする現在のインターネットにおけるディレクトリサービスは人間の力により情報の分類、整理を行っているが現在でもすべての情報を網羅するには程遠いのが現状である。

DLA 法を用い DLA 情報の微分によりキーワードの頻度を解析し、ピーク位置の近似を計算行えば、キーワードのマッチングのみよりも精度がよく、言語処理よりも計算量の少ない自動的な文書の分類も可能である。

5 問題点

5.1 キーワードの選択

DLA 法ではキーワード選択の前に単語のマッチングを行い、不要な単語を除去している。一般的な動詞、助詞、助動詞や冠詞などがこのリストに所属する。現在はリストを自分で作成しているが、本来ならば既存の辞書などに頼るべきである。

また、キーワードとして選択した単語に関して活用形を考慮に入れていないため 単数形と複数形は別の単語として計算されている。これらの複数の形態を持つ一つの単語の取扱いに関する配慮が必要であり、複数形、活用形に関する辞書を持つことを考えている。

5.2 文書のサイズと分割数

電子化文書にはさまざまなサイズのものが存在する。WWW 上で 1 ページ分程度しか分量のないものから、従来の書籍をまるごと電子化したようなものまで幅広い。今回の実験では一般的な論文を対象として、文書を便宜上 10 分割し、ヒストグラムを作成した。

そのため文書サイズが違えば、一つの部分文書に含まれる単語の数が大幅に違うという問題が発生する。このことはヒストグラムの傾向を分析するには問題にならないが、ヒストグラムそのものを比較は行いにくい。

そこで文書を決められたサイズで分割し、その何番目にキーワードが存在するかという方法を行うプログラムとの比較を行う必要がある。

後者は紙のイメージで何枚目に情報が存在するかを示すものであり、より従来からの自然な感覚に近いものだと思われる。

しかしながら、どちらの方法もあまりにも小さな文書の場合には効果を発揮しない。これは分割数や分割後の文書に含まれる単語数が非常に少なくなるため、重要な語句の含まれる確率が低くなるからである。

5.3 複合語の取扱い

現在の実装では情報の最小単位を単語においている。しかしながら、いくつかの単語が組合わさった複合語の検索に対する要求も強い。しかしながら、複合語の検索を可能にするには単語の接続情報を採り入れる必要がある。既存の電子化辞書に頼るのが一番簡易な方法であるが、最新の話題に対する新しい造語を積極的に採り入れた電子化辞書を自由に利用できるようにはなっていない。

よって、ユーザが自分達の必要とする複合語を単語ではなく複合語として認識させるためのユーザ定義辞書の機能は必須であると考えられる。

5.4 日本語の対応

今回の実験では、英語の文書に関して DLA 法を適用した。日本語に対応させるためには、日本語文書からの単語の切り出しを行う必要があるが、これに関しては自然言語処理関連の成果を利用しなくてはならない。

日本語の場合には同じ単語でも漢字、カタカナ、ひらがな、アルファベットなど多種の表記方法が存在するため、これらの統一も課題の一つである。特に漢字とひらがな、カタカナの相互変換や、送りがなの揺れの問題は辞書なしには解決できないため自由に研究に利用できる機械可読な辞書が望まれる。

6 今後の課題

今後はこのシステムを電子図書館のデータに適用し、一般ユーザを対象とした実証実験を行うことを予定している。

現在の実装では地図と呼んでいる画像を検索者が見ることにより、判断を行っているが、グラフの特徴を抽出することにより、自動的に検索や分類を行えるシステムを構築したいと考えている。

7 おわりに

本研究では、キーワードの位置情報に着目し、位置と頻度を集計することにより、従来のキーワードマッチングでは得られないような文書の内容に関する情報を、簡単な手法で得ることができることを示した。

しかしながら、単語の活用の取扱い、複合語の取扱い、プログラムによる類似度の判断などの問題を残しており、これらの改善を行う必要がある。

また、今後はもっと幅広いデータに対し DLA 法を適用し、広範囲なデータをもとに検討を進めたい。

参考文献

[1] Marti A. Hearst. Tilebars. http://elib.cs.berkeley.edu/tilebars/about.html.

[2] Martin Gleeson. fly: create gif images on the fly. http://www.unimelb.edu.au/fly/.

[3] NSF/DARPA/NASA Digital Libraries Initiative. D-lib magazine. http://www.dlib.org.

[4] Yahoo. http://www.yahoo.com.