データベースシステム 120－3
（2000．1．24）

語の出現位置の視覚的記憶に基づく既読文書の問合せに於ける索引構造の考察

日高宗一郎 加藤弘之 大山敬三

学術情報センター

情報検索に於いて文書のレイアウトに関する視覚的記憶を頼りにした検索を実現するための、語の出現位置を考慮した検索方式について、その索引構造の考察を行う。我々は先に既読の文書に対する視覚的記憶に基づく検索の重要性を、情報の個人化の文脈の下に指摘し、実現の方向性を提案し性能に関する議論を行った。しかしそこでの索引構造は伝統的な転置ファイルの拡張であり、出現位置からの検索の高速化に配慮していなかった。本稿では多次元インデックスで文書のページレイアウト上の出現位置と検索語を扱うことにより、検索語と位置の両観点から高速に検索を行うための索引構造への展望を述べる。

A Study on Index Structure for Querying Already-Read Documents based on Visual Memory of Word Occurrences

Soichiro HIDAKA, Hiroyuki KATO and Keizo OYAMA

National Center for Science Information Systems(NACSIS)

Index structure concerning index term occurrence position to realize Information retrieval based on visual memory of index term occurrence in page layout is studied. We previously pointed out the importance of visual memory search on already-read documents in the context of information personalization, proposed implementation strategies, and discussed several performance aspects. However, since suggested indexing strategy was an extension of traditional inverted file, it could not be expected to perform well at searching based on occurrence position in page layout. This paper studies an alternative efficient indexing strategy to unify index term with occurrence position using multi-dimensional index structure.

1. はじめに

Web などの情報検索は通常読んだことのない文書を対象に行われる。しかし、受信した電子メール、自らの作成した論文、報告書等の、一度目を通したことのある文書に対する検索もまた重要である。既読文書の場合記憶にレイアウトの印象が残っていることが多いと考えられるため、視覚的記憶に基づく検索が行えることが望まれる。

我々は、先にこうした既読文書に対する視覚的記憶を頼りにした検索の重要性を、情報の個人化の文脈の下に指摘し、実現の方向性を提案し性能に関する議論を行った [3]。しかし、そこでの索引構造は伝統的な転置ファイルの拡張であり、出現位置からの検索の高速化に配慮していなかった。また、ユーザの検索入力と結果の出力について、実際の出現位置と入力との間にユーザの認める距離感覚についての議論も行われていなかった。

本稿では、多次元インデックスでページレイアウト上の出現位置と検索語を統一して扱うことにより辞書順中の語の位置とページレイアウト上の位置の両観点から高速に検索を行うための索引構造への展望を述べるとともに、検索入力と実際の語の出現位置との間の距離感覚に関する考察を行う。

以下、第 2.節で先に提案した視覚的記憶に基づく検索についての概略を述べ、次に第 3.節で単語情報と出現情報の多次元索引による統合について考察する。更に第 4.節では検索インタフェースを実際の語の出現位置との間の距離感覚に関して再考する。

2. 視覚的記憶に基づく検索

視覚的記憶に基づく検索の重要性については前節で述べた。この枠組を我々は通常の検索語に基づく検索と組み合わせて用いることを仮定している。本節では本論文で想定される検索インタフェースと、アーキテクチャの概要、転置索引の拡張に基づく語の出現位置を埋め込んだ索引方式について順に述べる。

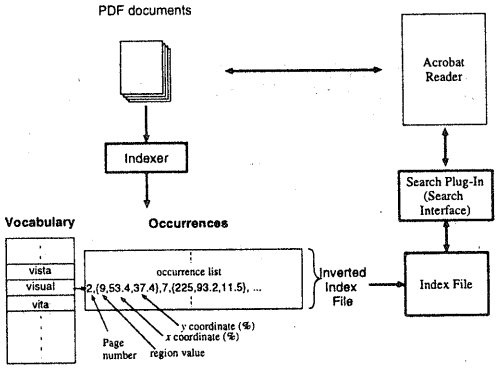

図 1 Implementation example for PDF documents

2.1 検索インタフェース

検索語毎にページレイアウト上での出現位置を指示出来るインタフェースを用意する。ウィンドウ上の矩形領域内の点およびそこに出現する検索語を指定する。検索結果は出現位置を考慮しない従来の適合度に出現位置による適合度を考慮したランキングにより表示される。

2.2 アーキテクチャの概要

提案システムは大別してレイアウトの確定した文書から出現位置を含む索引を構築する部分 (インデクサ) と、ユーザからの検索要求を処理し結果を提示する部分 (検索インタフェース) により成る。個人化を考慮して索引はユーザ毎に作成され、出現位置情報はレイアウトエンジン (個人化前の一時文書データベースからレイアウトの確定した個人化後の文書への変換を行う過程でレイアウトを決定するコンポーネント) から得る。

2.3 転置索引の拡張による出現位置の表現

ここでは索引構造として考えられる実装戦略について、先に提案した転置索引の拡張に基づく構造 [3] について述べる (図 1)。

図では、PDF 形式の文書から各単語のページ上の位置を取得して位置情報を含む転置索引を生成し、Acrobat Reader ブラウザのプラグインとして検索インタフェースを実現する方式を示している。Occurrence リストの各出現情報毎

に、ページレイアウト上の位置がx座標およびy座標の組で追加されている。Vocabulary部は主記憶に搭載可能と考えられるが、Occurrence部は文書サイズによりいくらでも大きくなるので主記憶には搭載不可能であるため配置等には注意が必要である。

3. 単語情報と出現情報の統合

転置ファイルの拡張方式では、文書の絞り込み時に単語の字句的情報からoccurrenceリストを特定し、出現位置による絞り込みは逐次検索になるため十分な性能性能が得られない可能性がある。ここでは、多次元索引にに単語の辞書順中の位置をひとつの次元として格納する索引方式について考察する。

3.1 多次元索引

多次元索引手法については既に様々なデータ構造が提案されている[2, 4]が、扱うデータの性質を基にどのデータ構造を用いるか、もしくは専用の構造を新たに準備するか見極めなければならない。

本論文で扱う格納データは各軸について以下の特徴を持つ(ここでは値の特徴、(もし分かれば)典型値、値の数の典型値について列挙する)。

- 語彙軸: 転置索引に於けるVocabulary部[1]と同等の性質を持つ。
- 文書ID軸: 整数
- ページ軸: 整数
- リージョン値軸(フレーズ検索の都合上必要となる場合を想定している)
- X軸: 実数、連続的 値の数は10〜20、値の間の距離は2文字分以上
- Y軸: 実数、離散的 値数40位(行数に等しい)、値の間の距離は行間に相当するため文書毎に共通

次元の数は全部で6になる(従来の多次元索引方式に於いて次元数が10を上回る辺りから逐次検索に比べて性能が劣化するという報告があるが[5]、この場合は10より小さいため問題とならない)。

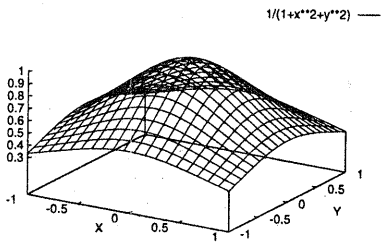

図2 ユークリッド距離

これ等の情報を用いて索引(ファイル)として最適化すべきパラメータがブロックの大きさ、分割の手法を検討することが出来る。

4. 検索インタフェースの再考

我々はこれまでユーザの出現位置の検索入力を点によるものと仮定していたが、明示的に何らかの大きさを持った領域を指定したい場合も考えられる。また、検索対象の検索語の出現とユーザの入力との類似度について、ユーザの認める類似度も単純にユークリッド距離で測定して良いのか議論の余地が残されている。物理的なレイアウト上の距離とユーザの認める距離が一致するとは限らないからである。

一般には、位置に関してユーザの認める類似度は座標の関数として表現することが出来る。図2、3、4にそれぞれ検索入力点からのユークリッド距離、マンハッタン距離、その他ピラミッド型の関数形の場合を例示する。ピラミッド型の関数形は、文書の文字の配置がY軸に関して一定間隔に規則的に配列されていることおよびページが矩形であることを考慮したものである。

5. 結論

索引構造について具体的な構築と検索手法について考察した。本稿執筆の時点では、既存の多次元索引のどれが本提案手法に適しているか、若しくは新しい索引構造が求められるか明

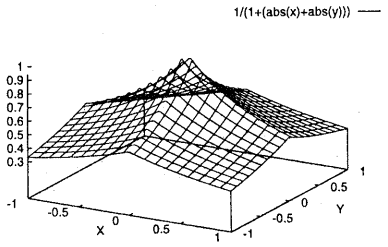

図 3 マンハッタン距離

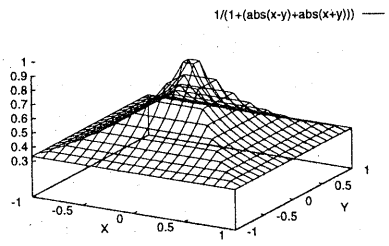

図 4 ピラミッド型距離関数

らかになっていない。

今後の課題 今後は索引構造を具体化する必要がある。その際、本稿で検討したユーザ毎のレイアウト上の距離の間隔を反映した索引構造はどうあるべきかを考慮する。

また、従来の索引語のみを考慮したランキングと出現位置によるランキングをどう組み合わせて用いるか検討する。

検索インタフェースに関しても、ユーザの望む理想的な visual memory 問合せはどうあるべきか更に考察し、点による指定だけでなく範囲による指定を支援する必要がある。

参考文献

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* ACM Press/Addison Wesley, 1999.

[2] Volker Gaede and Oliver Guenther. Multidimensional access methods. *ACM Computing Surveys*, Vol. 30, No. 2, pp. 170–231, 1998.

[3] Soichiro Hidaka, Hiroyuki Kato, and Keizo Oyama. Querying Structured Documents based on Human Visual Memory. In *Proceedings of the International Symposium on Database Applications in Non-Traditional Environments*, pp. 226–229, Kyoto, November 1999.

[4] N. Katayama and S. Satoh. The SR-tree: An index structure for high-dimensional nearest neighbor queries. In *Proceedings of the ACM SIGMOD Int. Conf. on Management of Data*, pp. 369–380, Tucson, Arizon, 1997.

[5] Roger Weber, Hans-J. Scheck, and Stephen Blott. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *Proceedings of the 24th VLDB Conference*, pp. 194–205, New York, 1998.