ディジタル図書館のための概念情報を用いた科学技術論文の検索

堀井 千夏 今井 正和 千原 國宏
奈良先端科学技術大学院大学
〒 630-01 奈良県生駒市高山町 8916-5
Tel: 07437-2-5205, Fax: 07437-2-5379
E-mail: {chinatsu, imai, chinatsu}@is.aist-nara.ac.jp

概要

従来の図書館では, 図書館司書により, 検索指示語に対する概念の個人差を解消し, 情報の信頼性を保ってきた. しかし, ディジタル図書館では, 人間が介在しないことを目指すため, 図書館司書に代わって, 検索結果の適合性問題を解消する自動検索システムが求められる. 現在の自動検索システムは, 同義語・異表記処理などの表層的な手法で適合率を向上させている. しかし, 単語体系に基づいた手法では, 検索質問の意味解釈や利用者の検索意図の不足情報に対する処理が不十分であり, より深層的な観点からの検索が必要とされる.

本手法では,EDR 電子化辞書を用いて, 概念レベルでの検索手法を提案し, 検索質問と文献中の記述語を単なる文字列としてではなく, 意味的な情報として検索に用いる. 検索対象には, 現在, 本学のディジタル図書館で, 誤字・脱字を含んだ OCR 結果として保管されている科学技術論文を用いる. 科学技術論文には, 必ず主題が存在し, その主題ついて集中的に論じられているため, 記述語がもつ概念から, 概念レベルでの論文の主題を獲得し, 検索に用いる.

キーワード

ディジタル図書館, 概念情報, 検索質問, 索引語, 検索式, 科学技術論文

An Information Retrieval using Conceptual Index Term for Technical Paper on Digital Library

Chinatsu HORII, Masakazu IMAI and Kunihiro CHIHARA
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara 630-01, Japan
Tel: 07437-2-5205, Fax: 07437-2-5379
E-mail: {chinatsu, imai, chinatsu}@is.aist-nara.ac.jp

Abstract

This paper presents an approach for semantic information retrieval which is implemented on Digital Library. It is well known that Digital Library make the information retrieval automatic and possible to access immediately the every kind of media information from anywhere. However, no improvement is made for use of Digital Library about the retrieval errors based on individual differences of the concepts or senses of user's request. This is one of the significant problem for searching efficiency of information retrieval. The approach developed in this study uses not user's request itself but also concepts of the user's request to solve this problem. This makes possible to do the semantic information retrieval not merely to do the comparison of the word strings of the user's request.

keywords

Digital Library, Concept path, Query, Index Term, Request, Technical Paper

1. はじめに

近年におけるネットワークやマルチメディア技術の著しい発達にともない, 情報提供の窓口となる「ディジタル図書館」の実現を期待する声が高まっている. 本学においても, 平成8年4月よりディジタル図書館の運営が開始され, 研究が進められている. ディジタル図書館は, 従来の図書館の機能に加え, 映像や音声等といった様々なメディアに情報源の区別なくアクセスすることが可能であり, 情報の即時性や, 時間や場所に対する拘束がないことが大きな利点である. しかしながら, 情報の電子化が進むにつれて情報量が膨大となり, 利用者の要求に対する適切な情報の提供がディジタル図書館を実現するうえで重要な鍵となる.

従来の図書館では, 検索質問に対する概念の個人差から生じる検索モレやノイズを図書館司書が解消し, 情報の信頼性を保ってきた. しかしながら, ディジタル図書館では人間の介在を減少させるため, 図書館司書に変わって検索結果の適合性問題を解消する自動検索システムが求められる.

これまでに, 情報検索システムとして, 検索式の表現や索引語の選定に関する様々な手法が提案されてきた. 検索式の表現法には, 論理演算に基づいた Boolean query language が広く用いられている. この手法は検索式の表現力を向上させるが, 利用者から入力された検索質問の文字列にだけ着目しているため, 検索モレが生じやすいといった欠点をもつ. そのため検索質問を意味的に拡張する手法として, 用語間の同義・階層・関連関係を表記したシソーラスが利用されている. シソーラスは同じような語の微妙なニュアンスの違いや含意の違いなどを使い分けるための辞書であり, ある語を手がかりとしてそれと同義語の関係にある語を探すことが可能である. しかし, 語を拡張する範囲の決定が困難であり求める意味ではない語まで含んでしまうといった問題を抱えており [1][2][3], 拡張範囲を絞り込み必要な語だけを選定する手法が求められる.

索引語の選定法には, 語句・語幹の類似性を数量的に解析する方法として確率論やベクトル空間, ファジィ集合論に基づいた研究がさかんに行なわれている [3][4][5]. しかし, 索引語を単語の出現頻度に基づいて決定しているため, 内容的特質を十分に表現しているとはいえない. そこで, 出現頻度に依存しない手法として, 文献の主題をグラフで表現する研究などが行なわれている. この手法は, 検索質問や記述語を単なる語の羅列ではなく, 構造的に内容のつながりを表現する手法である. 最適なグラフ表現や照合法が確立していないため, 照合に時間がかかることが問題点である [3].

本手法では, 検索式の表現および索引語の選定における上記の問題点を解消するために, 以下の特徴をもつ検索手法を提案する.

- 概念体系の階層構造に基づいた概念空間の意味的な絞り込み
- 概念の出現頻度の利用
- 概念情報に基づいた索引語の選定
- 概念の ID 番号化による照合の簡易化

本システムは従来の単語体系に基いた表層的な検索手法ではなく, 概念体系に基づいた深層的な観点からの検索手法である. シソーラスとして EDR 電子化辞書 [6] を用い, 検索式と文献中の記述語を単なる文字列としてではなく, 意味的な情報として検索に用いる. そのため, 語のうらに隠された意味や, 利用者による検索意図の不足情報を推測することが可能であり, 図書館司書の協力なしに利用者が真に求める検索結果を提供することが期待される. また, 電子化辞書を多言語シソーラスとして用いることにより, 自然言語処理分野で問題とされている異なる言語間における検索への対応が可能となる.

本研究では, 検索対象に科学技術論文を用いる. これは現在, ディジタル図書館で保管されている誤字・脱字を含んだ OCR 結果である. 科学技術論文には必ず主題が存在し, その主題ついて集中的に論じられて

いることに着目し, 文献中の記述語から概念レベルでの論文の主題を獲得することを目指す. このことは, 論文の位置付けやクラスタリングといった文献分類の実現が期待され, ディジタル図書館および情報検索分野において大きく貢献すると考えられる.

2. 概念空間の絞り込みによる概念の獲得

本研究では, 深層的な観点から, 情報検索を実現するために, 検索式および文献中の記述語を単なる文字列としてではなく, 意味的な情報として取り扱う. このことにより, 検索質問の意味解釈や利用者の検索意図における不足情報の推測が可能となる. 語の意味的な情報としては, 概念情報を用い、概念の階層構造に基づいた比較により概念空間を絞り込む.

2.1. 電子化辞書による概念情報

本手法では, 単語の概念情報を獲得するために, "EDR 電子化辞書" とよばれる大規模な機械処理用辞書を用いる. 従来の辞書は人手で構築されたものであり, 辞書にのっていない単語が比較的多く, 常に専門用語や新しい語を追加する必要があった. しかし, 最近では専門用語辞書の作成や, 国語辞書に基づいた概念体系（シソーラス）の構築に関する研究が行われており, 辞書情報を有効に使用することは効率のよい情報検索を可能にすると考えられる.

EDR 電子化辞書は 11 のサブ辞書から成る. 各サブ辞書は記述の単位（e.g. 語, 句, 文, 文章）, 言語の種類 (e.g. 英語, 日本語) と記述のレベル（e.g. 概念）の 3 つの特徴をもつ座標軸で構成されており, 統合的な言語知識を明らかにした辞書である [7]. 本手法ではサブ辞書の中から単語辞書, 概念体系辞書および概念見出し辞書を用いて, 語の概念情報を検索する.

2.2. 概念空間の絞り込み

概念体系辞書には, 概念間の関係として, 上位－下位関係が記されている. 語の上位概念を繰り返し検索することにより, 概念空間を木構造として体系化することができる. 例えば, "ベース" という語から構成された概念空間の主要な枝葉は, Fig.1 のように示される. しかしながら, 1 語から構成された概念空間は, 利用者の意図や論文の主題とは異なる意味の方向へも情報空間を広げるといった問題点をもつ. そのため, 本手法では概念体系の階層構造に基づいた比較により, 概念空間から必要な概念だけを選定し, 概念空間を絞り込む. "ベース" という語に "楽器" という語を追加した場合, これらの比較から, Fig.1 に示すような意味的な重なりが得られる. この部分を, 絞り込まれた概念情報とする. 各概念とその連結情報を概念パスと定義し, 概念パスの上位方向を概念空間の意味的な方向とする. 実際, 各概念情報は Fig.2 のように 16 進数のＩＤ番号で表示され, 検索を容易に行なうことが可能である.

3. 情報検索システムの概要

概念情報に基づいた情報検索システムは, 概念レベルで検索式を表現する検索式部, 科学技術論文の主題を選定する索引語作成部, および検索式と索引語を照合する照合部の 3 部から成る. 以下に, それぞれについて述べ, Fig.3 に本システムの概要を示す. 日本語の形態素解析には「JUMAN システム」[8] を用いる.

3.1. 検索式部における検索式の概念獲得

検索時の適合性問題は, 利用者が入力する語の妥当性に依存する. そのため, 検索の精度を上げるには, 語のうらに隠された意味や利用者の意図における不足情報を推測する必要がある. 本手法では, 語がもつ概念情報を用いて, 深層的な観点から検索式を導く. 以下に概念レベルで検索式を獲得する流れを簡単に述べる.

1. 利用者から入力された検索質問を ID 番号に変換する

2.ID 番号別に上位概念を検索し, 各概念体系を構成する

3. 概念空間を絞りこむ

4. 絞り込まれた概念パスの意味を利用者に提示する

絞り込まれた概念パスの中から利用者の意図にそったものを選択してもらい, 最終的な検索式とする.

3.2. 索引語作成部における科学技術論文の概念獲得

検索対象の文献には科学技術論文を用いる. 科学技術論文には必ず主題が存在し, その主題について集中的に論じられている. そのため, 論文の記述語から得られる概念パスを用いて, 概念レベルでの主題を獲得する. 以下に概念レベルで論文の主題を獲得する流れを簡単に述べる.

1. 論文のページ毎に OCR をかけ, テキストを獲得する

2. テキストを形態素解析し, 名詞の抽出および出現頻度を算定する

3. 各名詞を ID 番号に変換し, 各概念体系を構成する

4. 概念空間を絞りこむ

5. 概念パスの出現頻度を算定する

6. 名詞および概念パスの出現頻度より概念を重み付ける

7. 論文の主題を概念パスの集合として保管する

概念の重み関数は, 式 (1) のように設定する.

$$(\text{概念の重み}) = \frac{1}{2}\left(\frac{C_T}{max(C_T)} + \frac{C_S}{max(C_S)}\right) \qquad (1)$$

ただし, CT をページ毎に出現する単語 T の出現回数, max(CT) を最も多く出現した単語 CT の出現回数とし, さらに CS を単語 T から求まった概念パス S の出現回数, max(CS) を最も多く出現した概念パス CS の出現回数とする. 各ページ毎に, 概念パスを重み付けし, 論文単位で保管する. このことにより, 論文を概念の集合として求めることができ, 概念レベルでの論文の主題を獲得することが可能となる.

さらに, 論文間における適合率を比較するために, 式 (2) に示す指数を用いる.

$$(\text{指数}) = \frac{(\text{各概念パスの重み})}{(\text{論文のページ数})} \times 100 \qquad (2)$$

3.3. 概念照合部

概念で表現された検索式と, 概念の集合として保管されている科学技術論文とを照合し, その結果を WWW ブラウザを用いて, 利用者に提示する.

4. 実験と評価

4.1. 実験

本システムにおいては文献の主題を概念で表現することが焦点となる. そこで, 科学技術論文のテキストデータから論文の主題を概念パスの集合として獲得する実験を行なった. 以下に実験の条件を示す.

- 科学技術論文には, 本学のディジタル図書館が保有している OCR 結果を使用
- 上位概念をたどる回数は, 3 回
- 絞り込みに, しきい値は設けず, 一度しか出てこない概念パスのみを切り捨て
- SPARCstation-10, SunOS Release 4.1.3 を使用

Table 1 に, 電子情報通信学会論文誌に掲載されている" 初期視覚における網膜双曲細胞" に関する論文の実験結果を示す. これは主題の概念を重み順に並べたものであり, ID 番号とその意味を表している. 3 つずつのグループは, それぞれ概念パスを示しており, 下にいくに従って上位概念を意味する. 結果よりこの論文の主題は, 情報分野における視覚および通信（回路）といった概念の集合で表現されることがわかる.

4.2. 評価

実験結果を評価するために, 情報科学に関する知識をもつ 5 人に, アンケート調査を行なった. アンケートは, まず, 電子情報通信学会論文誌 VOL.J78-DIINO.7 に掲載されている 15 本の論文を読んでもらい, 『概念で表現した論文の主題が妥当であるかどうか』という問いに対して, "excellent" , "good", "fair", "poor" の 4 段階で評価してらうものである. アンケートの結果を Fig.4 に示す. その結果, 76 %の被験者が "excellent" または, "good" と評価しており, 概念で表現された論文の主題が被験者が想定する主題とほぼ一致しているいえる. また, "excellent" を 3 点, "good" を 2 点, "fair" を 1 点, "poor" を 0 点とし, 各論文の評価を点数化したグラフを Fig.5 に示す. 各論文に対する評価の平均点は, 15 点満点中 10.6 点と, 総合的には高いものであった. しかし, 論文の主題によっては, 非常に低い評価がなされた. このことは, 以下にあげる 2 点が問題であると考えられる.

- 電子化辞書に存在しない専門用語
- 形態素解析における不的確な連結規則

EDR 電子化辞書には専門用語の辞書が存在するが, 本システムでは使用しなかったため専門用語の概念化が十分に妥当ではなかったと思われる. 今後は, 専門用語辞書を検索システムに加える計画であり, さらに高い評価が期待される. また, 形態素解析には既存の連接規則を用いた. そのため, 複合名詞として意味をもつ語を分割してしまう場合が生じた. Juman システムにおける連接規則は, あらかじめ変更しておくことが可能であるので, 不的確な場合を十分に調べその変更を試みる必要がある.

5. 実装モデル

本研究で提案した検索システムを試験的に実装した. ユーザインターフェースには WWW ブラウザを用いる. 利用者による検索条件の入力から利用者が最終的に閲覧する論文のドキュメントにいたるまでの流れを順に, Fig.6, Fig.7, Fig.8, Fig.9 に示す.

6. おわりに

本研究では, ディジタル図書館のための概念情報に基づいた検索システムを提案した. 検索対象には, 本学のディジタル図書館で保管されている科学技術論文として, 誤字・脱字を含んだ OCR 結果を用いた.

以下に, 本システムの特徴を示す.

- 検索質問のあいまい性や冗長性の考慮
- 利用者の意図や論文の主題にそった概念空間の絞り込み
- 異なる言語間での検索

本手法では, EDR 電子化辞書を用いて検索式と科学技術論文の主題を概念で獲得し, 語を単なる文字列としてではなく, 意味的な情報として扱った. このことは, 表層的ではない, 深層的な観点における情報検索の実現を可能とする. また, 概念で科学技術論文の主題を表現することは, 論文の位置付けやクラスタリングの実現も期待され, ディジタル図書館および情報検索分野においても, 大きく貢献すると考えられる.

参考文献

[1] 諸橋正幸, 堤泰治郎, 丸山宏, 野美山浩: 情報検索システムにおける効果的ナビゲーション機能の提案. ディジタル図書館ワークショップ 第 2 回, pp.45-49.

[2]Lim, C. and Chen, H. "An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual(Chinese-English)Documents." IEEE Transactions on Systems, Man and Cybernetics, 1994.

[3] 細野公男偏: 情報検索. 雄山閣出版, 1991.

[4]Salton, G. Allen, J. and Buckley, C. "Automatic structuring and retrieval of large text files." Communications of the ACM 37.2 (1994). 94-108.

[5]Harter, S. P. "A probabilistic approach to automatic keyword indexing. Part II: An algorithm for probabilistic indexing." Journal of the American Society for Information Science 26.5 (1975). 280-289.

[6] 日本電子化辞書研究所: EDR 電子化辞書技術ガイド (第 2 版再改訂). EDR TR-045, 1995. [http://www.iijnet.or.jp/edr]

[7] 横井俊夫, 木村和広ほか: 表層レベルにおける電子化辞書の情報構造. 情報処理学会論文誌 Vol.37 No.3, pp.333-343.

[8] 松本裕治, 黒橋禎夫ほか: 日本語形態素解析システム JUMAN 使用説明書 version2.0. 奈良先端科学技術大学院大学

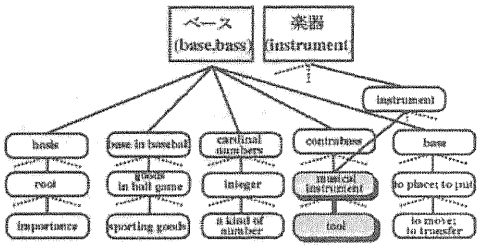

Fig.1. Narrowing down the concept space

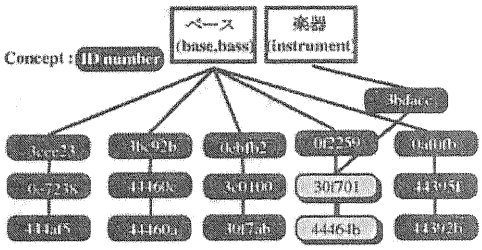

Fig.2: The concept space represented by the ID number

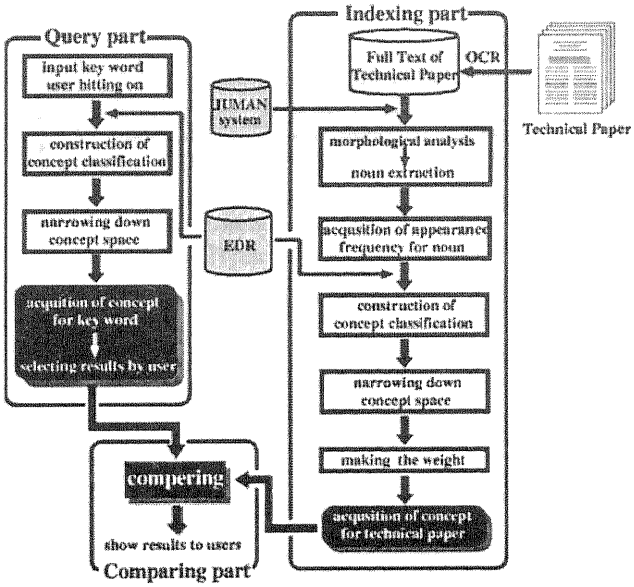

Fig.3: The overview of the IR system in the concept level

Table 1. A set of the concept path represented the theme of the technical paper

ID	Sense of ID	Weight	Weight Index
3c1947	retina (a part of an eye)	2.94	29
0f40b4	a visual organ, in animals	2.94	29
3d0f15	a sense organ (a type of physical organ that transmits received stimuli to nerves)	2.94	29
0e5230	reply (replication)	2.46	25
2f327b	speak (transmission of information)	2.46	25
2f324f	transmission of information	2.46	25
3ca8dd	bipolar cell	2.29	22
0f2e4c	cell (the smallest living unit whose aggregate forms the body of an organism)	2.29	22
3f961e	parts and element of the living things	2.29	22
3be671	circuit (a circuit for an electric current)	2.11	21
3aa92b	parts of a machine	2.11	21
2f2b32	parts	2.11	21

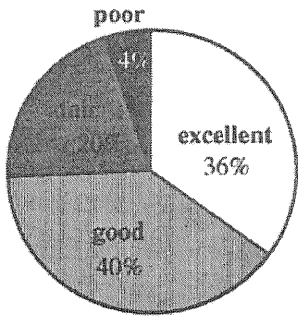

Fig.4. The result of the questionnaire

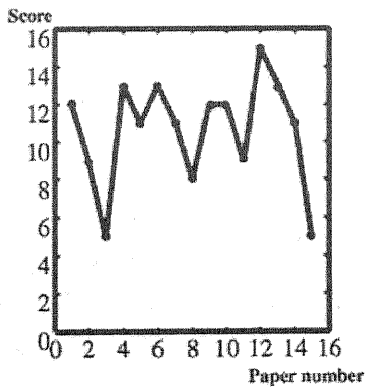

Fig.5: The evaluation results for each paper in Journal

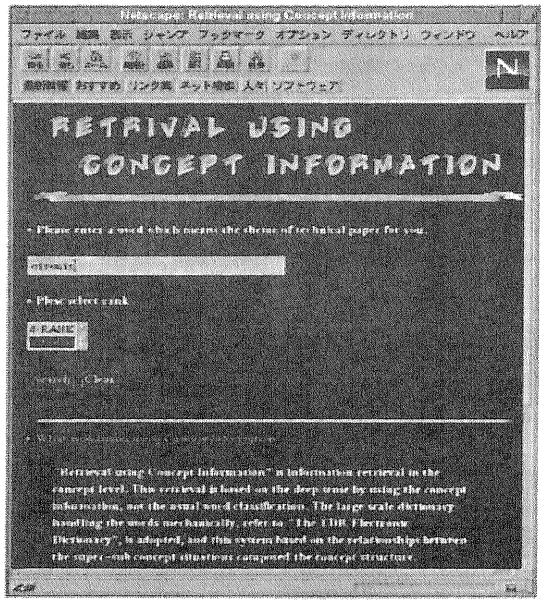

Fig.6. The retrieval form

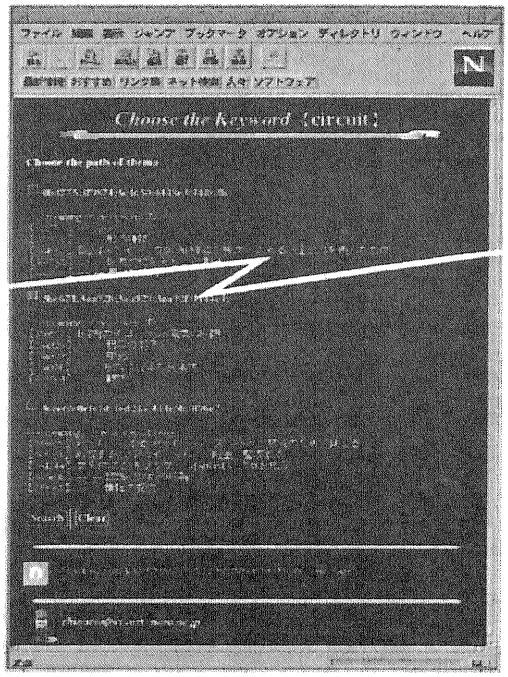

Fig.7. The list of the concept path

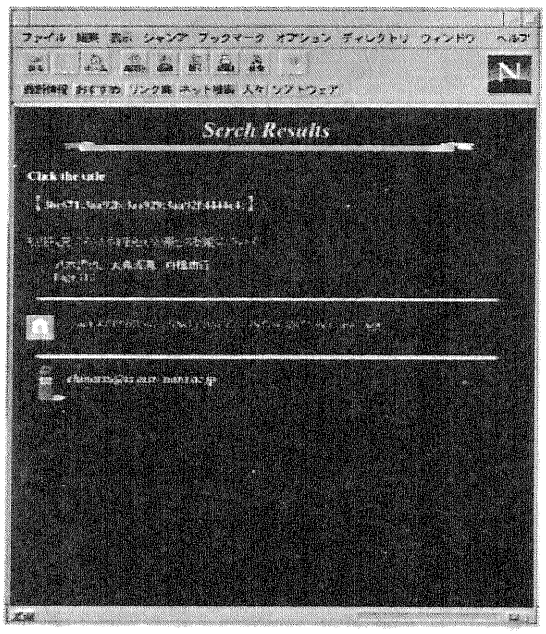

Fig.8. The title lists

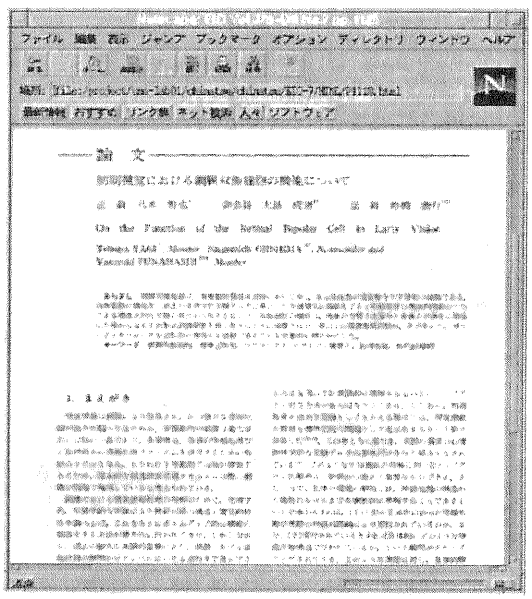

Fig.9: The document of the retrieval result