情 報 メ デ ィ ア 6－2
（1992．5．15）

情報通有におけるコンテンツと検索

（株）富士通研究所
松井 くにお　　吉岡 誠

1992年5月15日

概要

パソコン通信に代表されるネットワークシステムの普及で、大量の情報の交換共有が現実になってきている。このような傾向に対してマルチメディア・ハイパーメディアなどの各種のメディア及びメディアコンセプトが広がりつつある。そこで、真の意味での情報共有＝通有のためにはいかなる情報構造でなければならないか、またそのような情報構造を用いていかにして使いやすい情報検索インタフェイスを構築するかは極めて重要な課題である。

本報告では、このような情報通有のコンテンツアーキテクチャとしてSGMLをベースとすることを提案する。また、それらのデータの検索における再現率や適合率の向上のために有効なキーワード展開や、ユーザに対して適切な関連キーワードを提示するユーザインタフェースモデルを提案する。

The contens for the information sharing and its retrieval

FUJITSU LABORATORIES LTD.
Kunio MATSUI　　Makoto YOSHIOKA

abstract

Nowadays, we can easily share electronic documents through international communication networks. According to these trends, a lot of medias, like a multi-media or a hyper-media, and its concepts are expanding at the current situation. So, we have to consider about the true documents structure and its useful interface for retrieving.

In this paper, we propose the SGML based documents as the contents architecture for the information sharing. And we also report the useful user interface model for the information retrieval.

1　はじめに

パソコン通信に代表されるネットワークシステムの普及で、大量の情報の交換共有（これを公文俊平『ネットワーク社会』では情報通有と言っている）が現実になってきている。このような傾向に対してマルチメディア・ハイパーメディアなどの各種のメディア及びメディアコンセプトが広がりつつある。そこで、真の意味での情報共有＝通有のためにはいかなる情報構造でなければならないか、またそのような情報構造を用いていかにして使いやすい情報検索インタフェイスを構築するかは極めて重要な課題である。

さて、近年ＳＧＭＬの利用が日本でも話題になってきている。米国ではすでに公用文書の提出をＳＧＭＬのある決まった形式でなければ受け付けてもらえないほどのメジャーな地位を確保している。文書の流通における標準化という意味ではこのように大きな役割を果たしているが、その利用方法はこれからの課題である。今回は情報検索における実用的な観点から以下に報告する。

2　情報検索とＳＧＭＬ

情報検索におけるＳＧＭＬの利用として以下の２つの方法が考えられる。

1. 構造的なリンクを生かした情報のナビゲーション

2. タグに意味を持たせた情報の整理、発見

前者の方法では、今までの情報検索のパラダイムに捕らわれない新たな可能性を模索できるが、構造的なリンク付けの方法などに困難な問題も多く、大量の文書検索においては現在のところではあまり実用的ではない。そこで、今回は後者について考えを進めてみたい。これは、いわゆるＤＴＤに則って記述されたＳＧＭＬ文書のみならず、一定の項目に従って形式化された文書にも適用されるものである。ＳＧＭＬ文書の流通により、こういった文書の検索の必要性がより加速されることが考えられる。しかも単なるキーワード検索だけでなく、ユーザをサポートするべき有効な検索インタフェイスが必須となる。

3　情報検索の過程とインタフェイス

3.1　情報検索の過程

情報検索では言うまでもなく、ユーザの要求と検索対象のマッチングにおける過程が重要である。情報検索システムにおける検索の過程は、その検索の高精度化を目的とする観点と、検索対象の質的な変化を目的とする観点によって６個の部分に分けることができる（図１）。

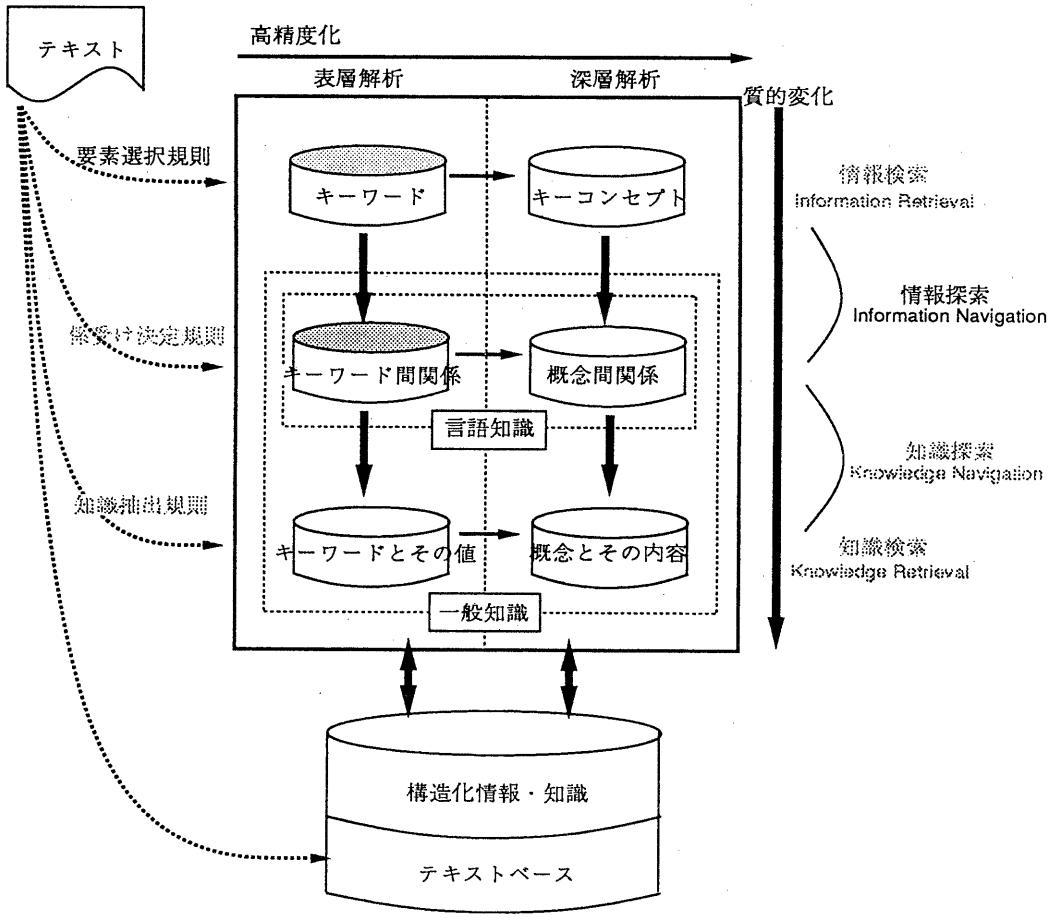

図1　情報検索の過程

高精度化という観点では、ユーザが検索システムに与えるトリガと検索対象の分析結果のレベルが一緒でなければならない。例えば、そのトリガがキーワードであるならば、検索対象は、人間が分析するか機械が分析するかは別として、キーワードがなければならない。概念レベルの検索としても、言葉の持つ表層に捕らわれない深層レベルの分析が必要であり [1]、ユーザがシステムに対して与えるトリガやシステムが保持すべき表現法は模索の段階である。

質的変化という観点では、キーワードやキーコンセプトという単発的なトリガに対して、そのキーワードやキーコンセプト間の関係をも含めた検索の方がユーザの求める検索対象にたどりつける可能性が高い。卑近な例としては、いわゆるAND、OR、NOTの論理演算では単語間の関係から検索対象を限定もしくは拡張している。しかしながらこういった絞り込みのための適切なキーワードをユーザ（専門に検索を行なうサーチャを除く）が知っている可能性は低く、システム側が知識として持ち適切なタイミングでユーザに選択を促すことが最も望ましいインタフェイスである。これらの知識は単語の類似性や共起性などの関係から、押し進めていけば機械翻訳などで用いている関係子までも含めた言語知識である。こういった知識を統計的にまとめあげることにより、キーワードもしくは、キーコンセプト間の関係を手繰りながらの情報探索が可能となる。また、その言語知識をさらに拡張し、一般の知識までもシステムが持つことができるようになると知識の検索が可能になる。ここでいう知識とはマッチングレベルの情報とは異なるもので、いわばSGMLのタグで示す中身の内容に相当する。これも言語知識を含めて考えれば知識探索が可能になる。ここでいう情報と知識の違いは、何らかの観点（SGMLのタグ等）でカテゴライズできるものを知識、そうでないものを情報と定義したい。

3.2 情報検索のインタフェイス

以前の本研究会で示した三輪は、情報検索システムにおけるユーザーインターフェイスの条件 [2] として以下の条件が論じている。

検索のステップは3段階まで メニュー方式の欠点として何段階ものステップを踏まなければ実際の検索にたどり着けないことを指摘している。経験的な感覚から3段階までが適当であるとしている。また、我々の調査 [3] によればキーワード検索における一般ユーザは、検索の限定や拡張を最も苦手としており、検索結果を全く得られずに検索を終了する場合も多い。それに対してサーチャは1～3段階の検索結果を得て検索を終了させる場合が多い。これは、検索自体のステップにおいても3段階までが適度な絞り込みの範囲であると言えよう。

自由キーワード方式では網羅的な検索は困難 同義語、類義語などのデータがないと、網羅的な検索ができないとしている。さらに、こういったものに加え、異表記の扱い、英語などの対訳語の扱い、入力されたキーワードを構成する語の扱いなども網羅的な検索を行なうためには必要なものと認識している。

外国語データベースの内容の翻訳 パソコン上での翻訳機能を望んでいる。パソコン通信やパソコンLANにおける機械翻訳もデータベースのスキャニング用としては、かなり充実してきたため、こういった利用法が確立しなければならない。

情報の読解・理解のスピードを高めることを支援するシステム 玉石混交の中から必要な情報を選び出すスピードは変わっていないと論じている。不用なものを除去したり、一つ一つの情報の要点を表示する仕組みが必要である。

検索結果の選択・分析・理解を支援する機能 情報の比較や、情報の品質判定機能などを論じている。この機能は情報の整理を対話的に行なえるようにすることが本質であり、それを視覚化する機能が必要である。整理している中から思いがけない情報の発見もこういった機能を充実させることによって実現可能である。

こういった情報検索システムのユーザの要求に対して、次章で論じるシステムの研究開発及び実験を行なっている。

4 システム構築の実現のための手段

4.1 システムの構成

情報探索システムを図2に示す。このシステムはユーザインタフェイス部、テキストベースシステム部、各種変換テーブル部から成る。

ユーザインタフェイス部

キーワードの入力・表示・選択を行ない、探索環境ファイルにはユーザ独自の環境（過去の検索履歴や変換テーブルの使用状況などのログ）を蓄え、ユーザが情報の探索を対話的に行なうことができるようにする。

テキストベースシステム部

キーワード自動抽出部では、テキストから形態素解析によってキーワードを切り出す。次に、切り出されたキーワードの品詞属性を見て必要な品詞のみを取り出す。さらに、不要語削除などの個別処理を施している。こうして取り出したキーワードは、その出現の頻度や他の語との距離（本来は係受け関係の距離が最も正確な値であるが現在は行なっていない）を計算し、関連性の高い語を関連語テーブルに登録する。

また、重要語抽出や構造化については次項以降に述べる。

各種変換テーブル

変換テーブルはOR展開に用いるものとAND展開に用いるものに大別できる。OR展開では、機械翻訳辞書から作成した対訳語テーブル、同義語・異表記テーブル、さらに、複合語を分解して作成した構成語テーブルから成る。この構成語は入力されたキーワードが複合語と判断できる場合に意味をなす範囲内で構成語分割を行ない、キーワードを展開するものである。

また、AND展開では、その目的としてはユーザが求めている要求を明確にするようなキーワード展開をサポートするか、もしくは、ユーザが気づいていないキーワードを知らせることにある。構成語から複合語への展開や、キーワード抽出の際に作成する関連語の展開のためのテーブルから成る。また、固有名詞約10,000語を700カテゴリに分類し、その分類項目名または、兄弟の項目の展開を行なえるようにしている。例えば「阪神」は『球団名』であり、「中日」や「広島」を導き出せるようなデータである。

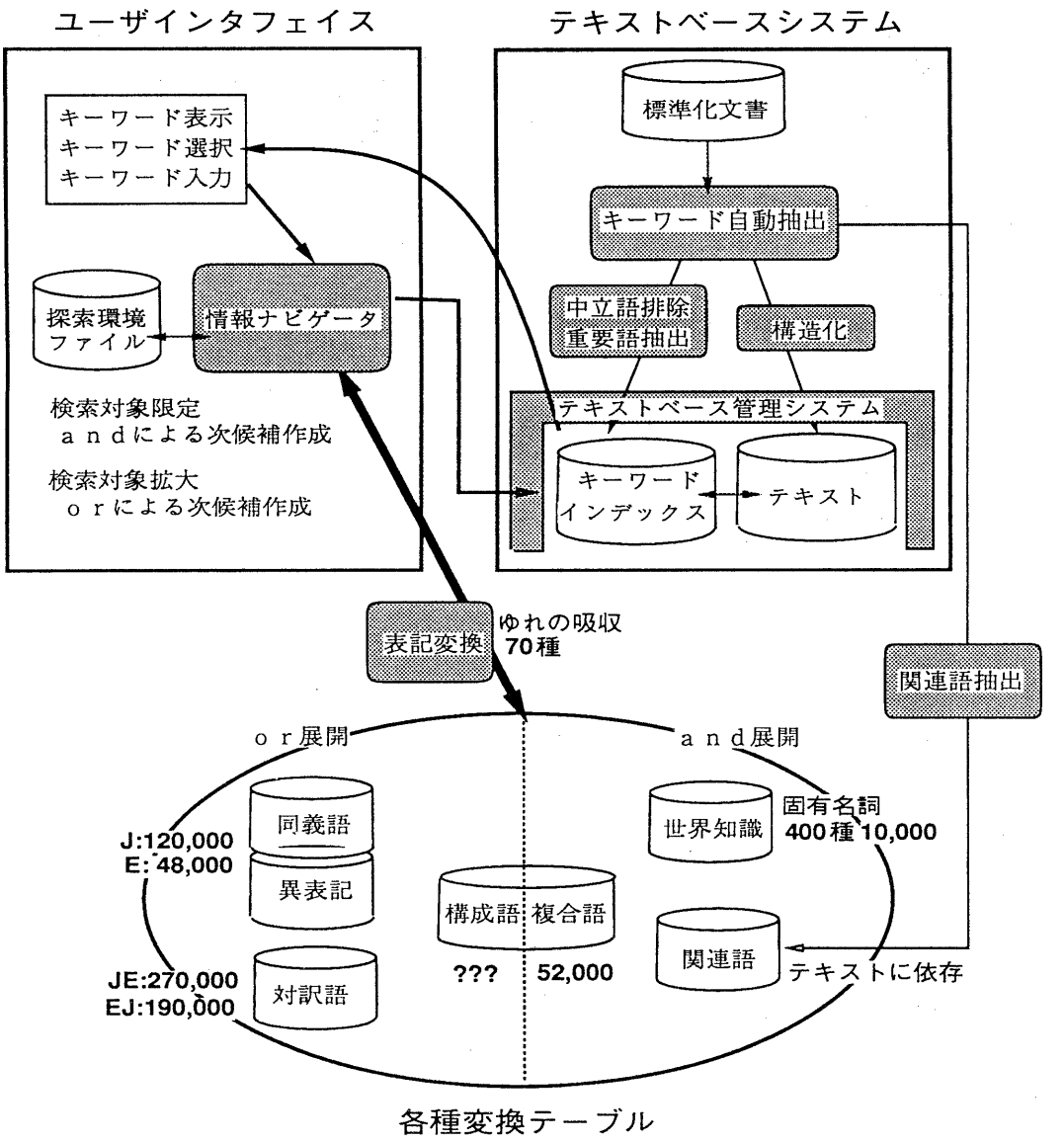

図2　情報探索システムの構成

4.2　キーワード展開

OR展開ではカタカナ表記などのゆれを吸収するための70種のパラメタを用意し、ユーザが任意に指定できるようにした。また、他の展開もユーザが個々に指定できるように環境ファイルで指定できるようにした。OR展開における表示は特に行なわないものとし、ユーザの要求があった時のみの表示にとどめる。

AND展開では図3に示すような関連語の展開とそれぞれの検索件数を先取りした表示を行なう。ユーザはこれによっておおよその検索件数を事前に知ることができる。複合語の展開や世界知識の展開もユーザの要求に応じて行なう。

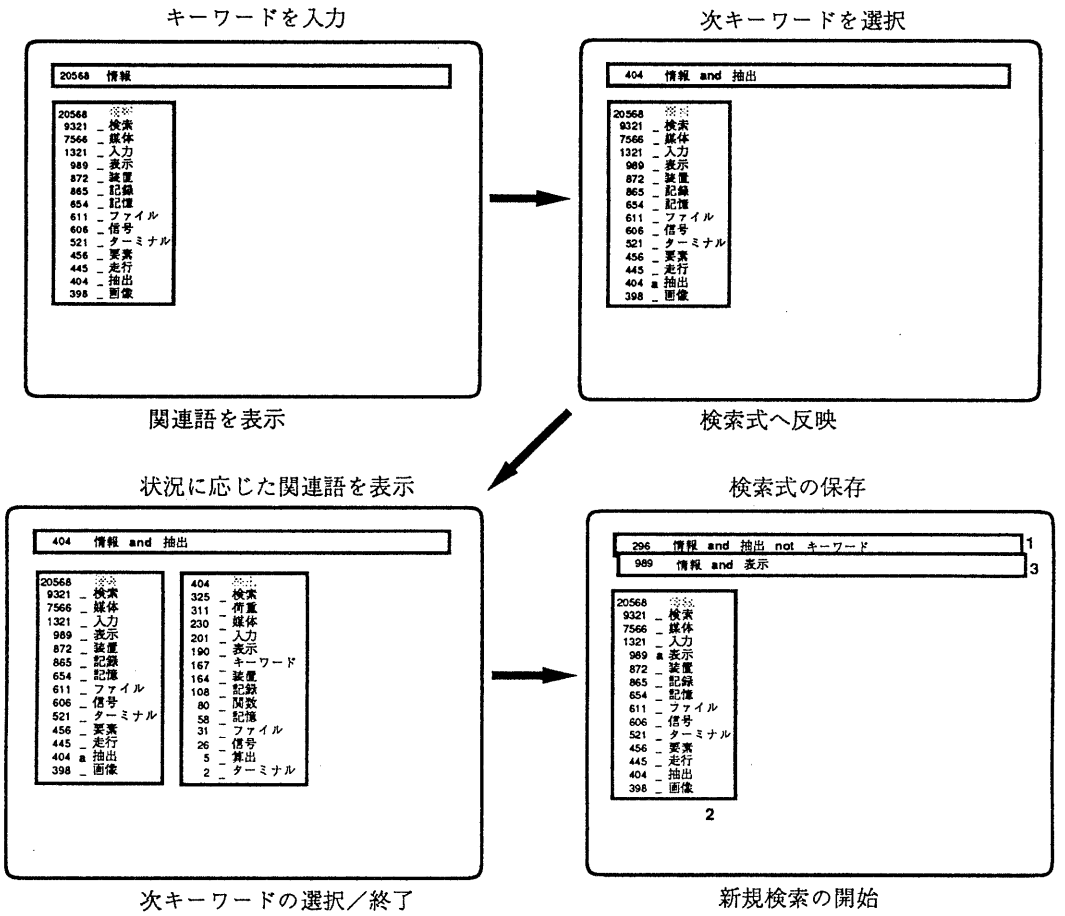

図3 関連語の展開

4.3 階層化されたテキストにおける重要語の抽出

パソコン通信などの元々階層化されたテキストにおいてはそのフォーラムの特徴を表すようなキーワードを導き出すことができる[4]。図4のような階層があった場合、最下位のレベルを個々に見れば特徴的であっても、横に並べると特徴ではなくなる各レベルでの中立語が存在する。例えば、音楽フォーラムの下位のフォーラムでは、以下の重要語が得られる。

演奏会フォーラム 「交響曲」、「ピアノ」、「協奏曲」、「ホール」、「席」、「曲」、「音楽」、「楽器」

吹奏楽フォーラム 「音楽」、「フルート」、「楽器」、「曲」、「笛」、「合奏」

合唱フォーラム 「曲」、「歌」、「音楽」、「ピアノ」、「楽譜」、「ハーモニー」

この場合、これらのフォーラムに共通するキーワードは「音楽」、「曲」であり、これらはレベル3のフォーラムでは中立語としてとらえることができる。しかしながら、レベル2ではこれらは最も音楽フォーラムの特徴を表すキーワードとなる。

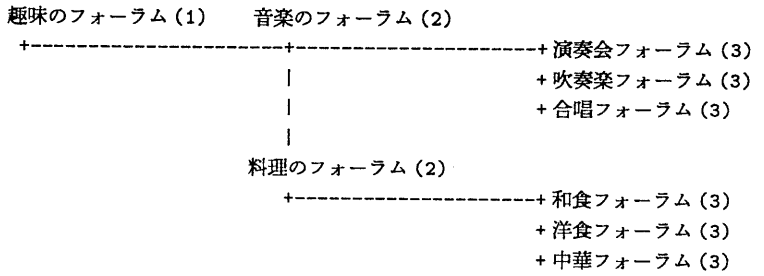

図4　階層化されたテキストにおける重要語

4.4　形式化されたテキストにおける情報の整理

形式化されたテキストのタグと関連語を利用すると図5に示すような情報を整理する表が簡単に作成することができる。これはタグの代わりに固有名詞の分類項目を使ってもよく、それぞれの分布が動的に作成できる。

部分集合：機械翻訳 **and** 言語
表の要素：関連語／〈出願人〉／件数

関連語＼出願人	A社	B社	C社	D社	E社	F社	G社
表示	5	4	0	35	10	29	11
解析	10	12	11	1	65	5	1
生成	12	22	46	2	33	4	2
辞書	35	4	2	2	11	34	7
支援	3	23	0	22	2	2	1
・・・・・・							

図5　情報の整理

4.5　検索過程の指標

検索の過程で不安を持ちながら行なっていることは、検索対象の絞り込みが十分であるか否かである。関連語間の関連性を見ることによってある程度の指標が出せるのではないだろうか。関連語同士の関連性を多変量解析などの手法を用いてその分散度合を考慮すれば、その絞り込みの度合が計れる。例えば分布がバラバラであれば絞り込みの度合は不十分であり、数個のグループに分かれるならばあと一歩であり、一つのグループにまとまるならばその絞り込みは十分であると判断できる。

5　実験の経過及び結果

キーワード展開においては、すでに変換テーブルのデータ作成を終了し、実験の段階にある。同義語展開では、特許文の検索においては適合率を落さずに再現率を向上させる実験結果を得ている。また、関連語の関連度の求め方は文書や文、係受け関係なども考慮して試行錯誤を行なっている。さらに、こういったデータの計算・表示機能も基本部はできており、高速化や大容量化に適合していく予定である。

6　おわりに

現状の情報検索に対して、その検索内容に形式化されたテキストを導入した場合を考慮して有用な検索インタフェイスについて提案を行なった。こういった情報検索についての評価基準は明確でなく、実験システムを作成しても定量的な評価がくだしにくいため、今後はその評価基準も併せて考えていく必要がある。

【参考文献】

1. 秋山「テキスト情報の知的検索における諸問題」情報処理学会データベース・システム研究会、1988.3.
2. 三輪「情報検索システムにおけるユーザインタフェースの条件」、情報処理学会情報メディア研究会、1991.9.
3. 下山、富士、松井「サーチャのノウハウに見る検索インタフェース」、情報処理学会ヒューマンインタフェース研究会、1992.3.
4. 富士「自然言語文書からの特徴キーワード抽出」、情報処理学会第43回全国大会論文集, 1991.