

高速全文不完全一致検索システムの実現

高田 寛
新日本製鐵(株)
技術開発本部
エレクトロニクス研究所
電子システム研究部

当社エレクトロニクス研究所では人間の連想構造を文書検索に応用することで、事前に各文書のキーワード付けあるいは辞書登録を行なうことなく、あいまいな語句による不完全一致検索が可能な超高速、大規模全文検索システムを実現した。また、本システムは専用ハードウェアを必要とせず全てC言語のライブラリとして実現されているため、様々な応用システムの構築が可能であり、その応用例を紹介する。

REALIZATION OF A VERY FAST RETRIEVAL SYSTEM WITH FUZZY SEARCH CAPABILITY

Hiroshi Takada
Electronics Research Laboratories
Technical Development Bureau
Nippon Steel Corporation
5-10-1 Fuchinobe, Sagamihara, Kanagawa, 229, Japan

Nippon Steel Corporation Electronics Research Laboratories has realized a very fast large-scale text retrieval system with fuzzy search capability. The system does not need any prior document indexing and is an outcome of application of human association mechanism research to text retrieval technology. The system does not require any special hardware and is provided as a library of C functions. The library is integrable to various application software as shown below.

自由キーワード全文検索システム

概要

本研究所では人間の連想機構を文書検索に工学的に応用することで、あいまいな語句による不完全一致検索も可能な超高速、大規模全文書検索システムを実現した。また、本システムは全てC言語のライブラリとして実現されているため、様々な応用システムの構築が可能である。

特徴

- 大規模文書データに対する全文検索を超高速に実現
- あいまいな語句による検索が可能
- 文書情報構造を自動学習するためにキーワード付けなどの作業が不要
- C言語のライブラリとして実現されており任意の応用システムへの組み込みが可能

仕様

- ・ 検索時間：約2秒
10文字程度のキーワード
文書データ：約92000件 約90MB
SPARC Station 2 1.3G SCSI HD
- ・ 不完全一致検索
- ・ 学習時間：10MB/hour
- ・ 61個のANSI準拠C言語ライブラリ関数
(学習用：35 検索用：22 その他：4)
- ・ プラットフォーム
NSSUN,IBMR6000,HITACHI2050,UNISYS など

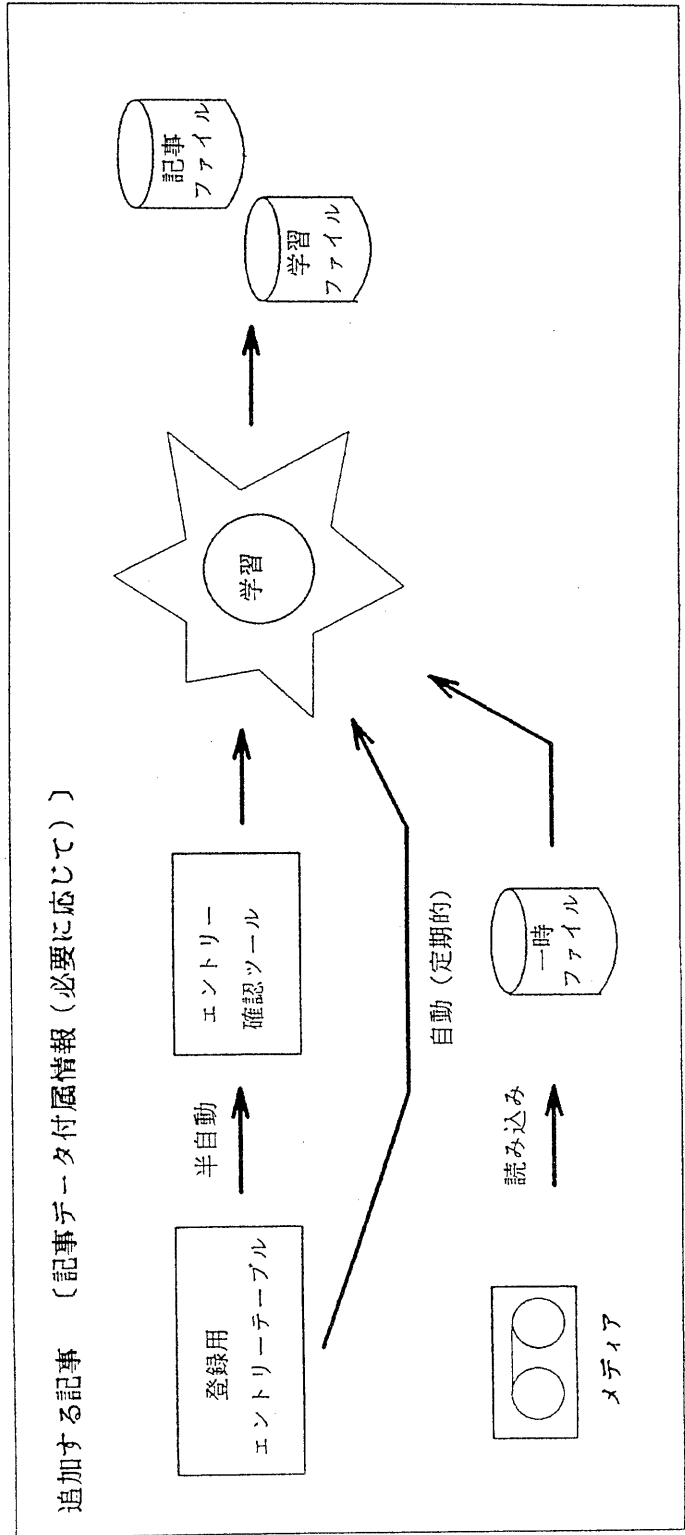
応用分野

- ・ 特許検索システム
- ・ 図書管理システム
- ・ 顧客管理システム
- ・ 百科事典検索システム
- ・ ソフトウェア管理システム
- ・ 技術情報管理システム

その他各種文書管理システム

学習

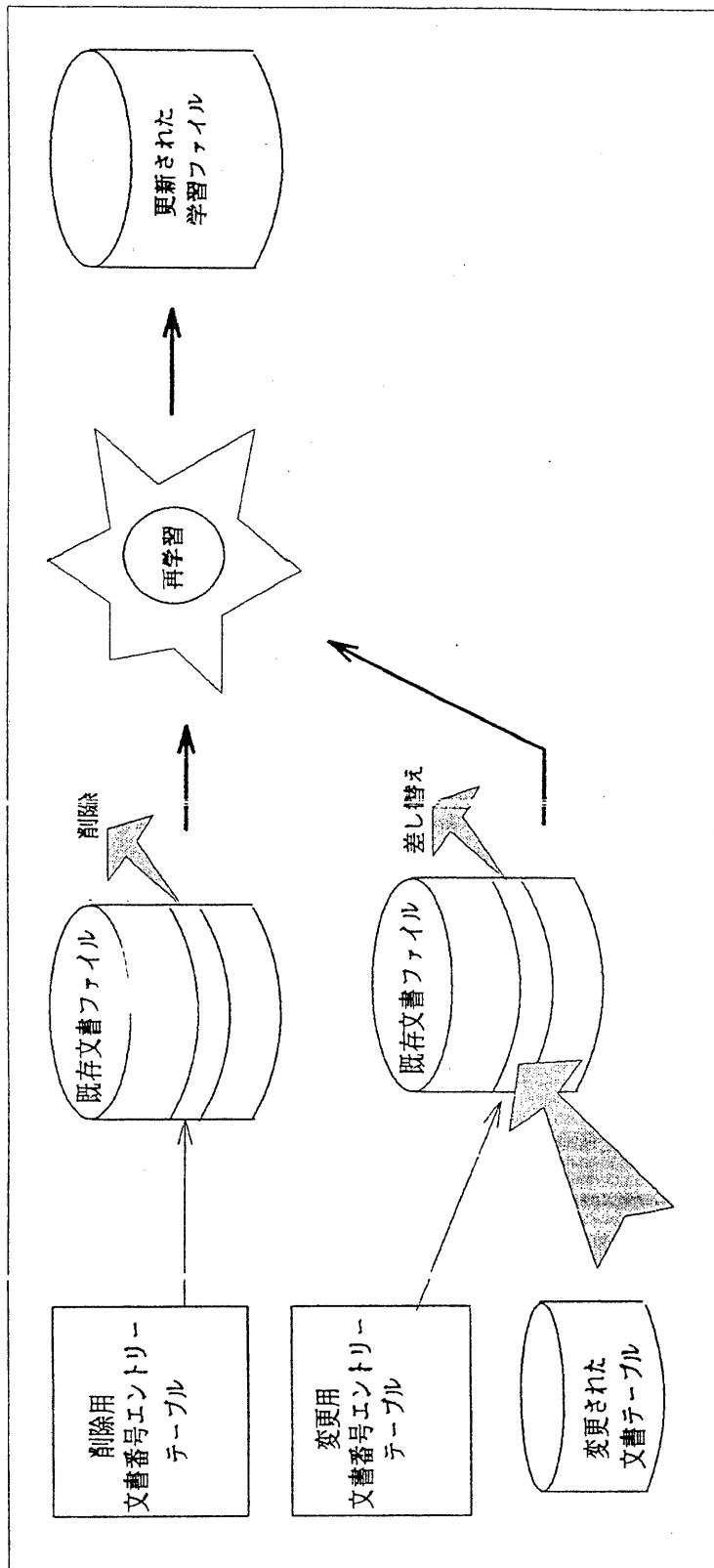
この検索システムは、学習という作業を経ることによって検索の対象とすることができるようになる。この学習は、従来のキーワードの切り出しなどは違って、システムがシステムのために行う作業のため、基本的にはユーザが介入する必要がありません。ユーザは、「いったい何を（どの文書/記事を）検索の対象とするのか」ということを決定すればよいだけです。学習では、システムが検索時に仕様する”学習ファイル”というものを作成し、記事を記事ファイルとしてまとめます。



文書メンテナンス

文書のメンテナンス機能として、文書の削除と変更が可能

- ・削除
 - レベル1：指定文書を検索対象から除外（削除文書の回復は可能）
 - レベル2：レベル1で検索対象から除外した文書の完全に消去
- ・変更
 - 指定文書データを異なる文書データに置き換えることが可能



外部システムとのリンク (応用例)

RDB、画像ファイルシステム、統合OAツール等と組み合わせることで
 トータルシステムの構築が可能です。

(例)

