

2N-1

情報検索システム構築に関する考察
— 検索手がかりの有効性の検討 —

平林 扶佐子

松浦 宏

日本電気㈱ C&C情報研究所

日本電気技術情報システム開発

1. はじめに

情報検索における検索精度を決定する主な要因に検索手がかりの構成と検索方式がある。検索対象となるドキュメントと検索要求の性質に適したものを選択しなければ高い検索精度は期待できない。

本稿では、検索手がかりの構成の検討指針について述べ、それに基づいた構成検討を支援するシステムを紹介する。

2. 検索手がかりの構成の重要性

情報検索システムにおいて十分な検索精度が得られない場合、まず始めに検索方式が原因ではないかと考えがちである。しかしながら、検索方式が適切であっても、検索手がかりの情報量が不十分であったり、構成が不適切であれば高い検索精度は期待できない。検索方式と検索手がかりの組み合わせが不適切な場合もある。そこで、検索方式だけでなく検索手がかりの構成の適切性も検討対象とする必要がある。

一般には、検索手がかりの種類が多い程、検索精度は向上する。従って、精度の向上の観点からは考えられる全ての検索手がかりを蓄積することが望ましい。しかし、実際には検索手がかりの作成コストが問題となるため、検索精度の向上と検索手がかりの作成コストのトレードオフを鑑みて、検索手がかりとしてどのような情報を蓄積するかを十分に検討しなければならない。

3. 検索手がかりの構成の検討指針

検索手がかりの構成の検討の第一の指針は、手がかり間の関係である。少ない手がかりで効率よく検索するためには、関連の低い項目を組み合わせることが必要である。

第二の指針は、検索手がかりとして利用できる確率である。識別力の高い手がかりであっても、まれにしか利用できないならば有効な手がかりとはなり得ない。

第三の指針は、重要な検索要求に対する有効性である。一般に、検索要求の種類により有効な手がかりは異なってくる。従って、各種の検索要求の発生頻度、重要度等を考慮して評価を行なう必要がある。

4. 支援ツール

3. で挙げた指針に基づいて検索手がかりの構成を検討するためには、大量の計算やシミュレーションを行うことが必要になる。そこで、支援ツールをパーソナルコンピュータ上で開発した。

索引方式は確率索引方式とし、ランク付け方式により一

つのドキュメントを検索する場合を対象とした。検索精度の評価尺度はサーチ可能枚数の制限の有無に応じて下記の2尺度を設定した。

- ① 目的とするドキュメントが上位x枚に含まれる確率
- ② 目的とするドキュメントを見つけるまでにサーチせねばならないドキュメント数の期待値

検索手がかりの構成を検討するための主な支援機能は以下の2つである。

a. 検索要求発生機能

ユーザによる直接指定あるいはシステムの自動発生機能により、種々の検索条件(検索項目値の組合せ)に対する精度評価を行う。

b. 検索結果の分析と評価機能

計算の途中結果、全体としての評価結果等をレベル分けして出力する。

検索要求の自動発生機能と評価機能を組み合わせることにより、発生し得る検索要求全体について、発生確率を重みとした検索精度の期待値を求めることなども可能である。

様々な構成について精度評価や検索シミュレーションが容易に行えるように、仮想データの利用、評価対象とする検索手がかりの構成の対話的な変更等の機能をサポートしている。

また、本ツールは汎用のリレーショナルデータベース上に構築されており、リレーショナルデータベース内に蓄積されたデータを直接用いることができる。従って、評価のための特別なデータ変換は不要である。また、システム内で通常のSQLによる検索等を行うことも可能であり、多面的な分析が行える。

5. おわりに

検索手がかりの構成の検討、精度評価における検索要求の重要度の考慮等は、従来の研究では見過ごされがちであった。しかしながら、検索システム構築の基本はこのような情報の分析であり、分析指針・方法を体系化し、ツールを揃えていくことが必要である。

[参考文献]

Shepherd, "An interactive computer system for retrieving faces" in Ellis et al.(Eds.), Aspects of face processing, Maltinus Nijhoff Publishers, Dordrecht, 1986