클러스터링을 이용한 시소러스 브라우저의 설계에 대한 이론적 연구

A Theoretical Study of Designing Thesaurus Browser by Clustering Algorithm

서 휘 (Whee Seo)*

〈목 차〉

초 록

본 연구는 전문정보를 대상으로 한 정보검색의 문제점이 정보탐색자의 탐색전략이나 기법에 대한 인식의 부족과 질의어 표현·생성·확장의 어려움에서 발생한다고 판단함에 의해 시작된다. 따라서 본 연구의 논제가 '시소러스 브라우저의 설계'에 관련된 내용임에도 불구하고, 연구된 내용은 내용 분석, 정보구조, 질의 형성, 질의 평가 등 자동정보검색의 전 분야를 망라하고 있다. 그 이유는 저자(연구자)들이 전문에서 사용하고 있는 용어와 탐색자의 질의용어를 일치시킬 수 있는 방법이 용어의 동시 출현빈도를 이용한 자동정보검색 방법이기 때문이다. 그러므로 본 논문의 연구방법은 선행이론 연구로써 이용자의 정보탐색 행태, 자동색인 작성법(automatic indexing), 클러스터링 기법, 시소러스 구축방법, 시소러스 구현방법, 정보검색기법에 대한 분석을 수행할 것이며, 그 결과로써 전문데이터베이스를 대상으로 한 정보검색의 효율 극대화를 위한 새로운 검색방법인 '클러스터링을 이용한 시소러스 브라우저' 구축의 이론적 모형을 제시할 것이다.

Abstract

This paper deals with the problems of information retrieval through full-text database which arise from both the deficiency of searching strategies or methods by information searcher and the difficulties of query representation, generation , extension, etc. In order to solve these problems, we should use automatic retrieval instead of manual retrieval in the past.

One of the ways to make the gap narrow between the terms by the writers and query by the searchers is that the query should be searched with the terms which the writers use. Thus, the preconditions which should be taken one accorded way to solve the problems are that all areas of information retrieval such as contents analysis, information structure, query formation, query evaluation, etc. should be solved as a coherence way.

We need to deal all the areas of automatic information retrieval for the efficiency of retrieval though this paper is trying to solve the design of thesaurus browser,

Thus, this paper shows the theoretical analyses about the form of information retrieval, automatic indexing, clustering technique, establishing and expressing thesaurus, and information retrieval technique. As the result of analyzing them, this paper shows us theoretical model, that is to say, the thesaurus browser by clustering algorithm. The result in the paper will be a theoretical basis on new retrieval algorithm.

* 창원전문대학 문헌정보과 조교수

Ⅰ. 서 론

지금 우리는 인터넷을 이용한 정보검색을 통해 필요한 정보 욕구를 해결하고 있다. 전통적인 서지 정보는 물론이고, 음성, 영상, 동화상 정보까지 해결할 수 있는 시대에 살고 있다. 더욱이 국내외의 전문데이터베이스는 원문을 대상으로 필요정보를 검색할 수 있도록 하고 있어 이용자들이 손쉽게 정보 욕구를 해결할 수 있다.

그러나 동일한 의미에 대해 저자와 탐색자가 동일한 형태의 언어를 사용하지 않고 있기 때문에 전문데이터베이스의 자연어 검색 효율이 저하되는 현상이 발생한다. 이 같은 현상은 저자가 사용한 용어를 탐색자가 정확히 알지 못하기 때문에 발생한다. 그 결과 데이터베이스 내에 원하는 정보가 분명히 수록되어 있음에도 불구하고 탐색자가 선택한 용어의 형태와 일치하지 못하여 해당 정보가 누락되는 현상을 초래한다. 이 같은 문제점을 해결하기 위해서 사전에 전문(fulltext)을 근거로 구축된 탐색용(질의 확장용) 시소러스의 필요성이 요구되고 있다.

그러나 해외의 경우와는 달리 국내에는 각 주제에 대한 시소러스가 거의 구축되어 있지 못해 이를 사용할 수 없는 형편이다. 또한 해당주제에 대한 한글 시소러스가 사전에 구축되어 있더라도 검색의 효율성이란 측면에서 상당히 의심스럽다는 것이 객관적인 의견일 것이다. 그 이유는 대부분의 기존 한글 시소러스가 해외 시소러스에 대한 단순 번역 과정에 의해 구축되었거나, 한글 전문에서의 어휘 출현 가능성에 대한 객관적 검증 없이 인간의 두뇌에 의한 어의 분석에 의존해 작성되었기 때문이다.

따라서 본 연구는 전문탐색에서 발생하는 이 같은 문제점이 정보탐색자의 탐색전략이나 기법에 대한 인식의 부족과 질의어 표현·생성·확장의 어려움에서 발생한다고 판단하고 이에 대한 해결을 모색하기 위해 시도되었다. 그 결과로써 초기질의어를 자동으로 분석해 성능이 좋은 질의어로 확장하고, 채택된 질의어를 바탕으로 자동으로 검색식을 구성해 탐색을 수행할 수 있는 방법에 대한 이론적 타당성과 함께 이를 근거한 검색시스템의 이론적 모형 구축을 목적으로 하고 있다.

그러므로 본 연구의 논제가 '시소러스 브라우저의 설계'임에도 불구하고, 이론적 연구내용은 일관성을 유지하기 위해서(이를 통해 검색의 효율성을 높이기 위해서) 자동정보검색(automatic information retrieval)의 모든 분야를 망라할 수밖에 없다. 그러므로 본 논문의 연구 내용은 이용자의 정보탐색 행태, 자동색인 작성법, 클러스터링 기법, 시소러스 구축방법, 시소러스 구현방법, 정보검색기법에 관련한 선행연구에 대한 이론적 분석과 함께 그 결과로써 전문데이터베이스를 대상으로 한 정보검색의 효율 극대화를 위한 이론적 모형, 즉 '클러스터링을 이용한 시소러스 브라우저' 구축을 위한 이론적 모형을 제시할 것이다.

Ⅱ. 시소러스 브라우저

1. 정의

정보검색시스템에서 온라인으로 시소러스를 이용하기 위해서는 시소러스에 표현된 용어 형태, 용어 사이의 관계 구조 등을 브라우징할 수 있는 시스템이 필요한데, 이를 시소러스 브라우저(Thesaurus Browser)라고 한다.[1] 시소러스 브라우저는 기존의 시소러스가 색인과 검색과정에서 병행해 사용되는 것과는 달리 온라인서비스에서 최종 이용자가 검색과정에서만 사용하기 위한 목적(자연어 시스템에서 어의적으로 관련된 용어나 동의어를 통제하기 위한 목적)으로 만들어졌기 때문에 탐색시소러스,[2] 자연어 시소러스,[3] 사후통제어휘집[4] 이라고도 불리워진다.

시소러스 브라우저는 기존의 시소러스와는 달리 탐색자가 용어선정을 표준화하기 위해 사용하는 것이 아니라, 탐색자 마음 속에 있는 용어의 대안어(동의어, 유사어, 반의어, 관련어)를 제공하는 기능[5]을 통해 주제탐색을 지원하는 정보검색 입장에서의 시소러스이므로 이용자에게서 발생하는 질문식과 문제기술의 불확실성, 문헌에 부여된 색인어의 다양성, 정보검색시스템의 복잡성을 해결하고 만족할만한 주제문헌을 찾도록 지원해 주는 것이다.

따라서 시소러스 브라우저는 탐색자가 단지 하나의 용어를 입력하더라도 이를 근거로 용어의 의미 네트워에 접근할 수 있도록 해야 하며, 탐색식을 형성하는 과정에서도 다양한 색인어를 적절히 조합한 다양한 탐색식을 제시하는 과정을 통해 자신의 정보 요구를 정확히 기술할 수 있는 기능을 제공해야 한다.[6] 또한 이용자의 용어 선택을 통해 문헌의 선택을 돕기 위해 문헌에서 추가적인 정보를 제공할 수 있어야 한다.

---

1) H. Albrechtsen, "PRESS : A thesaurus-based information system for software reuse", 《Proceedings of the International Study Conference on Classification Research》 vol. 5(1992), p.140.
2) W. Schmitz-Esser, "New approaches in thesaurus application", *International Classification*, vol. 18, no. 3 (1991), pp. 144-145.
3) F. W. Lancaster, 윤구호, 김태승 공역, 『정보검색시스템』. 서울 : 구미무역, 1985. pp. 314-315.
4) F. W. Lancaster, *Vocabulary Control for Information Retrieval*. 2nd ed., Virginia : Information Resources Press, 1986. pp. 165-169.
5) D. Soergel, *Organizing Information Principles of Database and Retrieval Systems*. New York : Academic Press, 1985. pp. 222-224.
6) M. Bates, "Subject access in online catalogs : A design model", *JASIS*, 37(1986), p. 366.

2. 형태

시소러스 브라우저의 형태는 표현 방법에 따라 평면시소러스 구조, 계층표시/트리 구조, 그래픽 구조, 순열 구조 등으로 나뉘어지며,[7] 한 화면에 2가지 이상의 구조로 표현하기도 한다.

어떠한 형태이든 특정개념에 관련된 모든 정보를 가능한 쉽고 신속하게 파악할 수 있는 기능을 제공해야 한다. 형태에 따른 시소러스 브라우저의 종류는 평면 시소러스, 계층표시·트리 구조 시소러스, 그래픽 시소러스, 순열표시 시소러스로 구분된다.

1) 평면 시소러스

한 화면 또는 윈도우를 단위로 참조 용어에 대한 범위주기, 관련어, 상하위 개념어 등의 정보가 2개 이상의 서브윈도우에 구분되어 표시되는 형식을 취하고 있으며, 구조 형태에 의해 평면 구조형태와 최상위개념어 구조형태로 구분된다.

평면구조형태는 인쇄형 시소러스와 거의 유사한 형식을 취하는데, 한 윈도우에는 선택된 용어의 범위주기, 관련어 등이 표시되고, 다른 윈도우에는 선택된 용어의 상위개념어, 하위개념어가 표시된다. 이 시소러스는 표시가 단순하다는 장점이 있는 반면, 상하위 개념어를 인덴션의 구별없이 평면적으로 나열하고 있어 출현 용어들의 관계를 즉각적으로 이해하기 힘들며, 계층구조의 파악이 힘들다는 단점이 있다. 대표적인 시소러스는 SilverPlatter의 WinSPIRS-ERIC이 해당된다.

최상위개념어 구조형태는 참조어에 대한 상하위 개념어와 함께 최상위개념어를 표시하고, 하위개념어는 해당되는 모든 디스크립터의 수를 부기하여 하위계층의 규모를 추측케 한다. 장점은 구조가 단순하여 이해하기 쉽고, 용어가 여러 계층에 속해 있을 때 각 상황을 이용자에게 보여줌으로써 이용자가 일차적인 탐색의 범위를 결정할 수 있는 장점이 있다. 대표적인 시소러스는 SilverPlatter의 WinSPIRS-MEDLINE이 해당된다.

이들 평면시소러스 구조는 관련계층을 파악하기 위해 계층관계에 있는 모든 디스크립터를 한번에 하나씩 찾아보아야 하며, 이용자가 이전에 살펴보았던 용어의 관계구조가 유지되지 않았기 때문에 전체적인 구조를 파악하고 기억하는데 어려움이 있다. 또한 범위주기, 용어 정의 등의 위치가 관련어의 앞에 있어 범위주기가 내용이 많은 경우 관련어가 한번에 파악되지 않는 불편함이 있다.

2) 계층표시/트리 구조

참조용어가 속한 전체 계층의 구조를 최상위 개념어부터 최하위 개념어까지를 나열하는 형

7) 이나니, 『시소러스 브라우저의 설계에 관한 연구』 석사학위논문, 이화여자대학교 대학원, 1996, pp. 22-29.

식으로 표시한 형태를 취한다. 표시 형태를 WinSPIRS-MEDLINE에서는 최상위 개념어를 '1'로 하고 각 계층에 계층 수준에 따라 숫자와 함께 '-'표기와 들여쓰기를 사용한다. 반면에 OVID의 MEDLINE에서는 접기기술(fold-in)을 이용하여 적은 공간에 효과적으로 계층을 표시하고 있다.

3) 그래픽 표시

시소러스는 하이퍼텍스트의 노드로, 용어 사이의 관계는 노드 사이의 링크로 표시한 그래픽 형태로 구조화되어 있다. 이 방법은 관련 개념간 용어 사이의 거리를 시각적인 그래프 형태로 연결시켜 제시해주므로 전체적인 개념관계를 파악하기 쉬우며, 전체 용어 사이의 개념관계를 등가, 계층, 연관 관계로 제한하지 않고 다양한 형태로 표시할 수 있어 의미구조의 시소러스에 적합하다.[8)]

4) 순열표시

가장 단순한 형태인 자모순 리스트에서 발전된 개념으로서 이용자가 입력한 단어가 시소러스에 있는지의 여부를 확인하고, 이 단어에 관련된 시소러스의 어휘나 어구를 자모순으로 제공해 주는 방법이다. 이 표시방법은 단어에 대한 순열색인을 제시해줄 뿐 아니라 복합어에 대해서도 가장 가까운 위치로 이용자를 안내해 줌으로써 단일어에 접근한 뒤 이를 포함하는 모든 어휘를 브로우징해야 하는 불편함을 줄여준다. 그러나 이 방법은 전후의 순열색인어가 한 화면에 나타나지 않아 어형변화 등에 따른 유용한 정보를 누락시킬 수 있다.

3. 필요성

1) 질의어 확장용 시스템

온라인정보서비스는 서지사항이나 초록 등의 2차 정보서비스에서 원정보인 전문 전체를 수록하고 있는 전문데이터베이스 서비스로 급속하게 변화하고 있다. 이같은 급속한 변화 발전의 이유는 컴퓨터 관련 설비 기술의 발전과 가격의 저하에 따라 컴퓨터가 보편화되고, 전반적인 인쇄·영상·음성매체 정보산업이 컴퓨터 지향으로 발전하고 있으며, 이용자들의 정보요구가 신속성을 보장하기 위해 온라인으로 1차정보원을 이용하려는 욕구가 증가하였기 때문이다.

---

8) E. B. Duncan, *A concept-map thesaurus as a knowledge-based hypertext interface to a bibliographic database*, London : Aslib, 1990. p.52

그러나 대부분의 전문데이터베이스가 사전에 색인작업을 수행하지 않기 때문에 탐색자에게 큰 부담으로 작용한다. 그 이유는 탐색식 구성시 용어의 조합을 적합문헌에 출현하는 정확한 용어를 예측해 조합해야 하지만, 특정주제에 대한 포괄적인 검색을 하기 위하여 제공되는 동의어, 계층어, 관련어를 예측하는 것은 저자의 저작유형이 다양하여 탐색자가 이러한 용어를 생각해내기는 무리이기 때문이다.[9]

특히 일반 이용자는 단일어(single word)만을 이용해 탐색을 수행하는 경향이 많기 때문에, 질의어에 포함된 용어가 적합문헌에 출현할 것이란 예측을 통해 검색작업을 수행하지만, 그 용어가 부적합문헌에도 출현할 가능성을 배제할 수 없기 때문에 예측했던 부적합문헌의 수보다 많은 부적합문헌이 검색될 수 있다.[10]

따라서 시소러스가 필요한데, 전통적인 시소러스는 수작업 색인(서지 데이터베이스 검색용 색인)용으로 작성되었기 때문에 자연언어색인을 기본으로 하고 있는(색인작업을 수행하지 않는) 전문데이터베이스의 검색에는 부적합하다. 그러므로 전문데이터베이스에서 검색효율을 높이기 위해서는 사전에 전문에서 출현빈도가 높은 용어들에 대해 동의어, 계층어, 관련어 등의 관계를 구축하여, 초기질의어에 대한 용어확장을 통해 검색을 수행할 수 있는 새로운 기능의 시소러스 브라우저가 필요하다.

2) 탐색전략 구축용 시스템 필요

대부분의 전문데이터베이스는 별도의 색인작업을 수행하지 않기 때문에 불필요한 정보의 출현과 필요한 정보의 누락이란 문제점이 발생한다. 그 원인은 이용자가 적용한 탐색용어의 부정확성, 용어 조합의 오류, 탐색 전략의 부적합성 때문이다.[11] 이같은 결과에 대해 Borgman(1996)은 최종 이용자가 특정 데이터베이스나 시스템에서 사용되는 시소러스나 주제명표목, 용어사전화일, 시스템 언어 등에 대한 이해 부족은 물론, 부울린 로직을 이용한 검색결과의 축소 및 확대에 필요한 탐색전략(strategy)과 기법(tactic)등에 대한 이해가 부족한데서 기인한다고 주장하고 있다.[12]

일반적으로 정보검색의 과정은 사전탐색(presearching), 데이터베이스 선택(DB selection), 탐색전략 구축(searching strategy construction), 온라인 탐색(online searching), 사후탐색(postsearching)의 단계를 거친다. 앞의 질의어 확장용 시스템은 사전탐색과 사후탐색에 해당되어 초기 질의어와 초기 검색결과를 이용해 계층어휘와 동의어 등의 관련 어휘를 추출하는

9) Lancaster, F. W., 윤구호, 김태승 공역, 『정보검색시스템』. 서울 : 구미무역, 1985. p. 313.
10) Moid A. Siddiqui, "Full-Text Database", *Online Review*, vol. 15, no. 6(1991), p. 369.
11) E. J. Mckinin. et al., "The Medline/full-text Tesearch Project", *JASIS*, vol. 42, no. 4(1991), p. 303.
12) C. L. Borgman, "Why are online catalogs still hard to use ?", *JASIS*, vol 47, no. 7(1996), pp. 493-503.

과정이다. 그러나 이 과정에 의해 추출된 확장된 어휘를 이용하더라도, 일반 이용자는 탐색전략 구축방법(and, or, not과 같은 부울린 로직을 적용시키는 방법)에 익숙하지 못하기 때문에 정확한 탐색식을 구축하지 못하고 있다.

따라서 탐색전략의 구축과 적용을 최종 이용자를 대신해서 수행할 수 있는 시스템이 필요하다. 이에 대한 가능성은 시소러스의 기능 확장에 대한 선행 연구결과에 의해 추측할 수 있다. Rowley(1994)는 시소러스의 기능 확장을 다음과 같이 설명하고 있다. "이용자 인터페이스는 GUI 또는 윈도우 환경에서 검색된 레코드나 탐색프로화일을 디스플레이 하는 동시에 스크린에서 시소러스를 찾거나 통제어휘의 리스트를 보는 것이 가능하다. 또 다른 발전은 자연어 탐색의 이용으로 지능 인터페이스를 제공하는 지식베이스로서 시소러스를 이용한다. 시소러스는 단어들 사이의 관계를 정의하고, 확장하거나 축소하는 등 다양한 방법으로 이용자 탐색을 발전시키는 시스템에 의해 자동적으로 이용할 수 있다"[13]

클러스터링을 이용한 시소러스 브라우저는 부울린 로직과 비부울린 로직(매칭함수에 의한 확률 검색방법)이 결합된 검색방법이므로 최종 이용자가 탐색전략과 기법등에 대한 이해 부족에서 발생하는 검색성능의 저하를 해결할 수 있다.

3) 이용자 지향적 시스템

현재 이용자 지향적인 수많은 시소러스 브라우저가 존재하고 있다. 이들 시소러스 브라우저들은 정보검색시스템이 DOS 환경에서 이용자에게 편리한 GUI 환경으로 바뀜에 따라 기존의 시소러스에서 사용할 수 없었던 다양한 관계기호들을 사용하여 색인자나 검색자에게 도움을 줄 수 있다. 그러나 대부분의 현재 이용되고 있는 시소러스 브라우저들은 수작업에 의해 작성된 시소러스를 단지 컴퓨터에 이식한 형태이므로 기능적인 측면에서 전통적인 시소러스가 갖는 기능에서 머물고 있다.

시소러스의 이용자 지향적 시스템의 가능성에 대하여 B. H. Weinberg(1995)는 시소러스의 혁신과 미래의 적용은 색인단계보다 탐색단계에서 확장이 가능하며, 특히 인공지능과 하이퍼텍스트 분야에서 필수적인 지식베이스의 구축에 시소러스가 기초가 되는 작업이라고 주장하고 있다.[14] 또한 Susan Jones(1995)는 계층구조와 같은 메뉴 인터페이스를 이용한 검색자와 문헌 간의 연결기능과 지능형 정보검색시스템을 제공하여 초기질의어 확장 기능을 갖춘 인간 중재 전문가시스템(human intermediary expert system)의 구축에 시소러스의 검색기능이 적

13) Rowley J. "The controlled versus natural indexing languages debate revisited : a perspective on information retrieval practice and research", *Journal of Information Science*, vol. 20, no.2(1994), pp. 115-116.

14) Weinberg B. H., "Library classification and information retrieval thesauri : comparison and contrast", *Cataloging and Classification Quarterly*, vol. 19, no.3/4(1995), p. 39.

용된다고 주장하고 있다.[15] 또한 최석두(1998)는 시소러스가 자연어 처리시스템의 기본사전으로서의 역할, 용어사전, 다국어 대역사전, 다국어 동적 시소러스로의 확장이 가능하다고 주장하고 있다.[16]

따라서 이용자 지향적 시스템의 성능을 갖춘 시소러스는 기존의 기능 뿐 아니라 온라인 환경에서의 정보검색의 효율성 제고를 위하여, 시스러스의 지식베이스로서의 가능성을 구현한 인간중재 전문가 시스템의 기능을 갖춰 사전탐색(presearching), 데이터베이스 선택(DB selection), 탐색전략 구축(searching strategy construction), 온라인 탐색(online searching), 사후탐색(postsearching) 등 정보검색과정의 전단계에 걸쳐 중요한 영향을 미치는 핵심적 시스템으로서의 기능을 갖추어야 한다.

이를 위하여 사전탐색과 사후탐색에 있어서 적합성을 근거한 초기 질의어의 자동확장, 데이터베이스 선택과 탐색전략 구축에 있어서 부울린 로직과 매칭함수를 결합한 탐색식 구성과 탐색 수행의 기능을 갖는 클러스터링을 이용한 시소러스 브라우저가 필요할 것이다.

4. 구축 과정

시소러스를 구축하는 방법은 기존 시소러스의 활용 여부와 어휘 선정시 주제전문가의 간섭 정도에 따라 구분된다. 기존 시소러스의 활용 여부에 따른 시소러스 구축방법은 다음과 같은 3가지 방법이 해당된다. 첫째는 같은 주제분야의 시소러스가 이미 만들어져 있는 경우, 최소한의 수정을 가한 후 그대로 사용하는 방법이며, 둘째는 기존의 일반적 시소러스나 관련 분야의 시소러스, 또는 주제명표·분류표 등의 어휘집을 전체적인 틀로 사용하되 핵심주제의 용어는 별도로 수집하여 상세한 시소러스를 개발하는 방법이고, 셋째는 새로이 용어를 수집하여 완전히 새로운 체제의 시소러스를 개발하는 방법이다.

주제전문가의 간섭 정도에 따른 시소러스 구축방법의 종류는 주제전문가들의 합의를 통하여 구축하는 방법과 문헌에서 용어를 추출하여 구축하는 방법이 있다. 이에 대한 많은 선행연구에서 주제전문가들의 노동집약적인 합의에 의한 방법보다 문헌에서 용어를 추출하여 구축한 시소러스로 검색한 결과가 최신용어의 수용면에서 효율이 높았다는 결과를 나타낸다. 또한 시소러스를 구축하는데 많은 인력과 비용과 시간적 노력이 소요되므로, 인력, 시간, 노력을 절감하고 시소러스를 용이하게 구축하고 최신성을 유지하는 자동 시소러스 구축방법에 대한 관

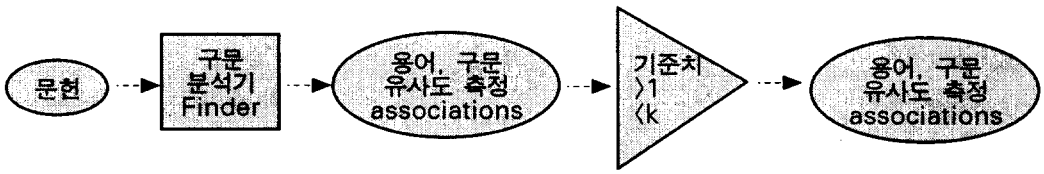
15) Susan Jones and others, "Interactive thesaurus navigation : intelligence rules OK?", *JASIS*, vol. 46, no. 1(1995), p. 52.

16) 최석두, "매크로시소러스에서의 용어 관리", 《전문용어언어공학센터 전문용어언어정보공학 심포지움》 제1권(1998), 1998, p. 44.

심이 높아지고 있다.

특히 시소러스에 출현하는 용어들은 문헌보증(literary warrant)과 이용자보증(user warrant) 원칙을 따라야 한다.[17] 문헌보증은 서지적 보증(bibliographic warrant) 이라고도 하며, 용어를 선정할 때 그 용어가 검색을 목적으로 중요하고 유용한 문헌에 충분히 출현했을 때에만 정당화될 수 있다는 의미이며, 이용자보증은 용어의 특정성 수준을 적절하게 구축할 때에 특히 중요한 것으로 문헌에 나타나는 용어는 이용자가 사용하는 것보다 더 특정적일 수 있기 때문에 필요하다.

따라서 본 논문에서는 시소러스의 문헌보증과 이용자보증 원칙에 따라 시소러스 구축방법을 선행 시소러스를 참고로 하지 않으며, 전문가들의 합의 과정이 아닌 반드시 문헌 내에 출현하는 용어만을 이용해 시소러스를 구축하는 방법에 대해 기술한다. 이같은 방법으로 시소러스를 구축하기 위해서는 <그림 1>과 같이 문헌내에서 핵심이 되는 용어들을 자동으로 추출하는 자동색인 방법이 필요하며, 이를 통해 빈번하게 함께 출현하는 용어쌍은 의미면에서 유사하거나 관련이 있을 것이란 가정에서 출발하는 용어의 동시출현빈도를 이용한 통계적인 방법인 클러스터링 방법에 대한 적용이 필수적이다. 용어 클러스터는 용어의 동시출현 여부를 통계로 처리한 유사도 매트릭스를 근거로 일정한 기준치 이상의 용어들을 동일한 개념으로 인식케 하는 방법인 퍼지이론을 적용하여 형성된다.



〈그림 1〉 시소러스 구축방법

앞에 설명한 시소러스 구축방법을 단계적으로 설명하면 다음과 같다.

첫 번째 단계는 어휘생성 단계로서 적절한 문헌을 수집하고, 시소러스의 특정성을 결정하고, 이를 근거로 문헌의 서명, 초록, 원문을 대상으로 용어를 수집한다. 수집된 용어를 근거로 정규화과정을 거치는데, 이 과정은 전치사, 접속사 같은 단어를 불용어 목록과 비교해 의미있는 어휘로 변형하고, 이를 어근형태로 변형시키는 스테밍 과정을 통해 색인어를 추출한다.

두 번째 단계는 어휘의 유사성 정도를 측정하는 단계로써, 먼저 문헌-용어행렬을 구성하고,

17) F. W. Lancaster, *Vocaulary Control for Information Retrieval*. 2nd ed. Virginia : Information Resources Press, 1986. pp. 23- 28.

이에 용어들의 동시출현 빈도를 고려한 cosine, dice같은 유사성 측정공식을 적용해 문헌간의 유사성을 확인한다. 이 단계에서 포함된 모든 문헌들은 계층화를 이루도록 클러스터링 알고리즘을 적용하며, 이때 클러스터의 표현은 클러스터를 형성케한 센트로이드에 해당하는 용어들로 표현한다.

세 번째 단계는 어휘의 관련도를 근거로 상위어, 하위어, 관련어로 표현하는 단계이다. 그 방법은 최상위레벨부터 시작해 인접레벨(adjacent level) 클러스터의 센트로이드를 비교해 처리하는 단계를 거친다. 먼저 직계계층인 부모-자식 형태로 이어지는 인접레벨(adjacent level) 클러스터의 센트로이드를 비교해 하위레벨 센트로이드에 새롭게 출현하는 용어는 하위어로, 상위레벨 센트로이드에 출현하는 용어는 상위어로 할당하며, 2단계 아래 인접레벨에서 형성된 클러스터를 비교해 공통용어가 아닌 용어는 상호간 관련어로 인식시켜 표현한다.

이상과 같은 단계를 거쳐 형성된 시소러스는 의미 네트웍, 지식 데이터베이스 등의 구축에 기초적인 대안이 될 것이다. 본 논문에서는 질의어 자동확장을 통한 검색효율성 확대를 주 논제로 다루었으므로 3번째 단계인 용어간의 관계 표현은 큰 의미가 없다. 그 이유는 이용자는 질의어 안에 상위어, 하위어, 관련어 등을 모두 포함해 정보를 검색할 것이며, 따라서 이용자의 정보검색 행태를 근거로 탐색용 시소러스 브라우저를 구축할 때 각 용어의 계층을 제시해 검색하는 방법보다, 클러스터 센트로이드에 해당하는 용어들을 묶어서 제시하는 것이 더 효과적이기 때문이다.

다음의 장들에서는 시소러스 브라우저를 구축하기 위해 필요한 각 단계에 대한 이론적 근거를 제시하기로 한다.

Ⅲ. 시소러스 브라우저 구축의 이론적 배경

1. 자동색인

1) 이론적 배경

데이터베이스에 저장되어 있는 수많은 정보에서 관련된 정보를 탐색하기 위해 문헌을 분석한 후, 주요 용어를 추출하여 수록된 정보와 연관시키는 작업을 색인이라고 한다. 색인의 종류는 색인어의 통제여부에 따라 통제언어색인과 자연언어색인으로 구분할 수 있으며, 컴퓨터의 간섭 정도에 따라 수작업색인, 반자동색인, 자동색인으로 나뉘어진다.

자동색인은 컴퓨터에 입력된 문헌을 대상으로 분석한 후 문헌의 내용을 나타낼 수 있는 단어나 단어구를 추출하는 과정이며 색인 과정에서 분석대상이 되는 부분은 문헌의 전문이나 초록이 된다. 문헌의 전문이 색인어 추출의 대상이 되는 경우에는 출현된 용어들에 대한 가중치를 부여해 색인어를 선정해야 한다. 이를 위해서 최소한 초록이나 본문 내의 서론, 결론을 대상으로 색인어를 선정하는 방법이 타당할 것이다.

컴퓨터에 의한 자동 색인은 시소러스 이용여부에 따라 시소러스 기반 색인법과 일반 색인 기법(단일어 색인 기법)으로 나뉘며, 일반 색인 기법은 색인어를 선정하는 기준에 따라 통계적 기법, 언어학적 기법, 문헌구조적 기법의 3가지로 나뉘어진다.

시소러스 기반 색인 기법은 연구자들에 따라 그 성능 평가가 다르다. 특히 국내의 시소러스는 문헌 내의 용어 출현 여부보다는 전문가들의 합의에 의한 방법과 해외 시소러스를 단순 번역해 구축한 것들이 대부분이므로 색인 용어의 불철저성(non-exhaustivity)이란 문제점을 안고 있으므로 이를 이용해 색인어를 추출하는 과정은 비합리적이다. 즉 잘못된 시작을 근거로 잘못된 결과를 발생케하는 문제점을 야기시킬 수 있으므로 본 논문에서는 색인작성법을 일반 색인 기법(단일어 색인 기법)으로 한정한다.

통계적 기법은 단어의 출현 빈도가 높을수록 그 단어가 문헌의 주제를 대표할 확률이 높다는 가설을 근거한 것으로서, 색인어 선정방법은 단어의 출현 빈도를 근거로 주제어로서의 중요도를 측정해 색인어를 선정한다.

언어학적 기법은 어휘적 단계, 구문적 단계, 어의적 단계로 나뉘며, 어휘적 단계 기법은 불용어 제거 기법을 의미하며, 구문적 단계 기법은 단어의 구문적 범주 결정을 위해 단어 사전을 사용하는 방법이 포함된다. 이 방법은 단서어 기법과 구문분석 기법이 해당되는데 그 중에서 구문분석 기법이 주류를 이루고 있으며 대부분의 구문분석 기법은 어의분석까지 포함하고 있다.

문헌구조적 기법은 문헌 속에 단어가 나타난 위치에 의해 색인어를 선정하는 기법으로서 서론, 본론, 요약 등의 제목을 갖는 특정한 부분에 나타난 주제들을 색인어로 선택하는 방법과 각 문단의 첫 문장과 마지막 문장과 같은 주제적 문장을 선택하여 이 문장 속에 나타난 주제어를 색인어로 선택하는 방법이 있다.

대부분의 한글 자동색인법은 언어학적 기법을 이용하여 색인의 대상이 되는 명사나 명사구를 식별하고, 통계적 기법을 이용하여 식별된 명사나 명사구를 색인어로 적용시키는 방법을 채택하고 있다.[18]

18) 남영준, 『색인어형태분석에 의한 한국어 자동색인기법 연구』. 박사학위논문, 중앙대학교 대학원, 1994.

2) 자동색인 알고리즘

자동으로 색인어를 추출하기 위해서는 어휘분석이 필요하다. 어휘 분석은 입력된 문자들을 의미있는 문자들로 변환하는 과정이다. 이를 통해 채택된 문자들은 후보색인 용어가 된다. 후보색인 용어들은 불용어 목록이나 불용어 사전과 대조되어 제거되며, 남은 문자들은 스테밍 과정을 통해 색인어로 변환된다. 각 단계를 분리해 설명하면 다음과 같다.[19]

(1) 어휘분석 알고리즘

문장 내에서 분리기호를 이용하여 어절을 분리하는 과정을 의미한다. 한글에서는 문자열과 문자열을 분리하는 식별자로써 공백문자(' ')를 사용한다. 그리고 생성된 문자열에서 <그림 2>에 제시된 기호들은 잘못된 문자로 인식하고 제외시킨다.

	.	!	?'	,	'	"	-	/	:	;	*	~	(	)	<	>	=	[	]	+	-	@

〈그림 2〉 어절 분리기호

나머지 문자열이 숫자로 시작하는 부분은 삭제를 한다. 단 숫자가 문자 다음에 공백없이 연결되는 것이나 중간점(·)을 통해 연결되는 것은 뒤의 문자와 함께 묶어 문자로 인식한다.
기호 중 하이픈(-), 대시(-), 마침표(.), 슬래쉬(/) 는 알파벳인 경우, 공백없이 뒤의 문자와 연결되는 경우 전체를 문자로 인식한다.

복합명사는 연결되어 표기될 경우에만 전체를 하나의 문자로 인식하며, 또한 용어 목록과 비교해 일치하는 부분은 단일명사 형태로 분리해 인식한다.

(2) 불용어 목록 구축 알고리즘

Luhn은 발생빈도가 높은 대부분의 용어들은 색인용어로 가치가 없음을 인식했다. 이들 용어를 사용하면 관련성과는 무관하게 데이터베이스의 모든 레코드를 검색하는 결과를 초래한다. 이들 단어들은 대부분의 문서에서 큰 비율을 차지하고 있다. 이들을 초기에 제거하면 검색속도와 성능의 향상과 색인 용량을 줄일 수 있다.

불용어를 추출하는 방법은 2가지 방법이 있다. 첫 번째는 어휘분석기를 통해 출력된 어절에서 불용어를 추출하는 방법이며, 두 번째는 어휘분석의 한 부분으로서 불용단어를 제거하는 방법이다. 첫 번째 방법은 모든 후보색인 용어들과 불용어목록을 대조해 보아야 한다는 문제

19) William B. Frakes & Ricardo Baeza-Yates, 류근호, 김진호 공역, 『정보검색』. 서울 : 시그마 프레스, 1994. pp. 154-205.

점이 있다. 이런 문제를 해결하는 가장 빠른 방법은 해싱기법을 적용하는 것이다.
두 번째 방법은 어절을 추출하는 어휘 분석과정에서 불용어목록과 대조하는 방법이다. 이 방법은 어휘 분석 과정 중 불용어목록을 갱신할 수 있다는 장점을 가지고 있다.

(3) 스테밍 알고리즘

스테밍은 색인화일의 크기를 축소하기 위해 적용된다. 용어대신 어간을 저장함으로 단일 어간을 복수의 완전 용어와 일치시키므로 50% 이상의 압축이 가능하다. 이 알고리즘은 접사제거 알고리즘, 후속자 변형 알고리즘, 테이블 탐색 알고리즘, n-gram 알고리즘이 있다.

접사제거 알고리즘은 하나의 어간을 남기기 위해 용어들의 접두사와 접미사를 제거하는 방법이다. 후속자 변형 알고리즘은 본문 내의 글자가 연속적으로 나타내는 빈도를 사용한다. n-gram방법은 용어가 공유할 수 있는 도표나 n-gram의 수에 기초한 용어들의 합성이다.

용어와 그에 상응한 어간은 하나의 테이블에 저장되며, 이 테이블을 확인함으로서 스테밍이 종료된다.

스테밍은 정확성, 검색효과, 압축성이란 측면에서 평가되어야 한다. 용어가 과도하게 어간화되면(과도 스테밍되면) 용어의 대부분이 제거되거나 무관련 용어가 합성될 수 있다. 이는 결국 관련이 없는 문헌이 검색되는 결과를 초래한다. 반면에 과소스테밍은 합성이 가능한 관련 용어의 결합을 어렵게 할 것이며, 그 결과 관련문헌이 검색되지 않는 결과를 초래한다.

한글인 경우 접사제거 알고리즘과 테이블 알고리즘을 이용하면 명사형 어근을 추출할 수 있을 것이다. 접사제거 알고리즘은 최장대응제거 방법과 단순제거 방법이 있다. 이를 위해서는 접사에 해당하는 용어들에 대한 테이블이 필요하다.

2. 클러스터링

1) 정의

현상세계는 수많은 존재들로 구성되어 있다. 이 존재들은 각 존재가 공유하고 있는 속성들의 크기에 의해 - 개체간의 유사성 차이에 의해[20] - 그룹으로 구성 또는 구분될 수 있다. 이 같이 어떤 대상들을 대상이 갖고있는 모든 속성들을 기준으로 일정하게 그룹화시키는 것을 분류(classification)라고 한다. 이러한 분류의 의미는 학문의 성향(분석대상의 차이)에 따라 classification, clustering, grouping, taxonomy 등의 다른 명칭으로 불리우고 있다.

---

20) Miyamoto S. and Nakayama K., "A technigue of two-stage clustering applied to environmental and civil engineering and related methods of citation analysis", *JASIS*, vol. 34, no.3(1983), p. 192.

클러스터링과 클러스터에 대한 선행논문의 다양한 정의[21][22]를 종합하면 다음과 같다. 클러스터링은 인간의 두뇌작업이 아닌 통계적인 빈도에 의해 대상간의 유사도를 측정하여 존재들을 유사하다고 판단하여 그룹화하는 방법이며, 여기서 생성된 그룹을 클러스터라고 한다.

클러스터를 형성하는 방법은 색인어-문헌 행렬을 이용하여 문헌 간의 유사치를 계산한 후, 이에 일정 기준치를 부여하여 기준치 이상 혹은 이하의 문헌들끼리만 연결시켜 문헌 클러스터를 구성한다. 형성된 클러스터에 공통으로 출현하는 용어들(센트로이드 벡터값에 해당하는 용어들)을 묶어 용어 클러스터를 형성하고, 이를 이용자의 질문식과 매칭시켜 해당되는 센트로이드를 제시해 질의어를 확장하는 방법으로 이용한다.

클러스터는 형성된 형태에 따라 스트링(string), 클럼프(clump 또는 connected components), 클리크(clique 또는 maximal complete subgraph)라고 불리운다. 스트링은 각 구성원이 서로 한번씩만 연결된 것을 의미하며, 클럼프는 각 구성원이 한번 이상 다른 구성원과 연결된 형태를 의미하며, 클리크는 각 대상이 서로 완전하게 연결된 형태를 의미한다. 또한 각 클러스터는 더 큰 클러스터의 최대 하부 그래프(maximal subgraph)라고 불리운다.[23][24][25]

2) 용어간 유사도 계산 방법

용어들간의 관계는 유사성(similarity), 연관성(association), 비유사성(dissimilarity)에 의해 측정된다. 비유사성은 용어 유사성을 측정하는 수학적방법에 해당되므로, 용어간의 관계를 표현하는 의미에 적합하지 않으며, 나머지 2개의 용어인 유사성과 연관성은 통계적 방법에 있어 동일한 의미이다. 일반적으로 클러스터링에서는 연관성(association)이란 용어를 더 많이 사용한다. 본 논문에서는 유사성, 연관성 그리고 유사도를 동일한 의미로 혼용해 사용하기로 한다.

용어간의 유사도 계산은 정보검색의 질의 확장시 용어 선정에 적용되며, 시소러스 구축시 용어간의 관계 설정시 필요하다. 용어간 유사도는 한 용어가 다른 용어에 얼마나 밀접한가를 의미하는데, 이는 전통적인 시소러스를 이용하는 방법과 corpus를 이용한 통계 정보를 이용하는 방법이 있다. 전통적 시소러스의 용어간 유사도는 단지 두 용어 사이에 몇 개의 용어가 연결되어 있는지를 통해 계산되며, 유사도 측정은 시소러스 구축자의 주관에 전적으로 의존된다. 반면에 corpus 통계정보를 이용하는 방법은 전통적 시소러스와는 달리 두 용어가 같이 출

21) Kendall M. G. and Buckland W. R., *Dictionary of Statistical Terms. 4th ed. London : International Institute*, 1982. p. 34.
22) T. Yu. Clement, "a clustering algorithm based on user queries," *JASIS*, vol. 25, no. 4, p. 219.
23) 이두영, "자동분류를 위한 document clustering의 기초 이론", 《중대 문리학보》 제39권, 1980, pp. 38-40
24) C. J. van Rijisbergen, *Information Retrieval*, 2nd ed. pp. 48-50
25) Miyamoto S. and Nakayama K., "A technigue of two-stage clustering applied to environmental and civil engineering and related methods of citation analysis", *JASIS*, vol. 34, no.3(1983), pp. 192-201.

현한 공기정보(collocation, co-occurrence information) 또는 상호정보(Mutual Information)등의 통계정보를 이용한다.

용어간 유사도를 계산하는 공식은 다음과 같다. 각 공식에서 X와 Y는 문헌을 의미하며, t는 각 문헌의 특성을 나타내는 용어들을 의미한다. 또한 Xt∩Yt 는 문헌 X와 Y에서의 공동출현 용어의 수를 의미하며, Xt∪Yt 는 전체출현 용어의 수를 의미한다.

① $|X_T \cap Y_T|$

② $2\times\frac{|X_T \cap Y_T|}{|X_T \cup Y_T|}$ ------- *Dice*공식

③ $|\frac{X_T \cap Y_T}{X_T \cup Y_T}|$ ------- *Jaccard*공식

④ $\frac{X_T \cap Y_T}{|X_T|^{\frac{1}{2}} \times |Y_T|^{\frac{1}{2}}}$ ------- *Cosine*공식

⑤ $\frac{|X_T \cap Y_T|}{Min(|X_T|,|Y_T|)}$ ------ 오버랩공식 (overlap)

⑥ $\frac{|X_T \cap Y_T|}{|X_T|+|X_T|-|X_T \cap Y_T|}$ ---------- *Tanimoto*공식

첫 번째 공식은 아주 간단한 공식으로 용어들이 일치하는 상관계수(coefficient)를 의미한다. Dice공식과 Jaccard 공식은 분모에 해당하는 계수를 동시출현한 용어를 각기 1로 계산한 것이고(동시출현 용어는 합쳐서 2가 됨), 오버랩 공식은 분모를 적은 어휘를 갖는 문헌의 어휘 수로 계산한 것이고, Tanimoto 공식은 분모를 동시 출현한 용어를 1로 계산하는 방법을 택하고 있다.

3) 클러스터링 알고리즘

클러스터링 알고리즘은 비계층적(Nonhierachical Clustering Algorithm)과 계층 클러스터링 알고리즘(Hierarchical Clustering Algorithm)으로 구분된다.[26] 비계층적 알고리즘은 용어간의 계층을 형성하지 않으므로 시소러스 브라우저 구축방법으로는 적합하지 않아 설명을 생략하기로 한다. 계층적 클러스터링 알고리즘(이하부터 계층 알고리즘)은 클러스터 대상물 간의 유사성을 측정하여 작성한 문헌-문헌 유사행렬을 이용하여 클러스터를 구성하는 방법이다.

26) Gerald Salton, *Dynamic Information and Library Processing*. New-jersey : Prentice-Hall, 1975. p. 329.

클러스터를 구성하는 일반적인 과정은 다음과 같다. 먼저 문헌-색인어 행렬을 대상으로 유사치(Similarity) 측정 공식[27][28]을 적용시켜 각 문헌쌍 { Di, Dj } (i = 1, 2,···, k / j = 1, 2,···, k) 들간의 유사계수를 계산해 문헌-문헌 유사계수 행렬을 구성하는 과정에서부터 시작한다. 여기에 일정 기준치(Trash-hold Value) 'T'를 부여해 Sim(Di, Dj) ≧ T 인 경우에는 '1'로, Sim(Di, Dj) < T 인 경우는 '0' 으로 하여 '1' 값을 갖는 문헌 쌍들에 소속된 문헌들을 하나의 클러스터에 소속토록 해, 서로 다른 문헌들을 동일한 문헌으로 간주하는 방법이다. 이 알고리즘에 의해 형성된 클러스터들은 유사도 순위에 따라 이원 나무 구조(binary tree struction)로 조직된다.

계층 알고리즘은 각 문헌이나 클러스터들이 모두 연결될 때까지 중복을 허용하는 방법으로 링크를 계속하는 방법을 택하므로, 비계층 알고리즘에 비해 공간(space)과 시간은 많이 요구되나 대상 문헌과 클러스터들이 계층을 형성케 되므로 문헌정보 검색에 더 적합한 알고리즘이다. 특히 시소러스는 의미에 대해 계층적 구조를 가지므로 시소러스 브라우저의 설계에 필요한 클러스터링 알고리즘은 분할 클러스터링 알고리즘보다는 클러스터들의 계층 구축이 가능한 계층적 알고리즘이 더 적절하다.

이 알고리즘은 클러스터를 형성하는 방법에 따라 응집적 방법(agglomerative)과 분열적(divisive) 방법이 있다. 응집적 방법은 클러스터가 이루어지지 않은 n개의 문헌 아이템에서 시작하여 n-1번의 결합이 이루어지며, 분열적 방법은 특정 클러스터에 소속된 모든 문헌 아이템들을 대상으로 n-1번의 결합을 통해 더 작은 클러스터들을 형성하는 과정을 거친다. 분열적인 방법은 거의 이용되지 않고 있어 유용한 알고리즘도 거의 존재하지 않으므로 생략하고 응집적 방법에 대해서만 논하기로 한다. 주로 많이 이용되는 응집적 방법은 단일연결(single link), 완전연결(complete link), 그룹평균연결(group average), Ward방법(Ward's method) 등이며, 기타 중앙값(Median)방법과 중심값(Centroid)방법이 있다.

4) 응집적 계층알고리즘의 종류

(1) 단일 연결 방법

단일연결 방법은 가장 널리 사용되는 방법으로 문헌간의 최대 유사도를 근거로 클러스터를 형성하며, 클러스터는 유사도가 가장 높은 문헌들부터 유사도가 낮은 문헌들 순으로 형성된다. 이 방법은 형성된 클러스터를 클러스터에 포함되지 않은 문헌들과 비교해 결합이 이루어지도록 하는 방법을 통해 새로운 클러스터를 구성한다.

---

27) Jardine N. and Rijisbergen C. J., "The use of hierachic clustering in information retrieval", *Information Storage and Retrieval*, vol. 7(1971), pp. 225-226.

28) van Rijisbergen C. J. *Information Retrieval*. 2nd ed. London : Butterworths, 1979. p. 39.

클러스터 간의 거리는 두 클러스터 중 하나에서 가장 가까운 점들에 해당되는 쌍(pairs) 사이의 거리로 정의되기 때문에 클러스터 중심(centroid)이 불필요하고, 처리하는 동안 유사행렬을 재계산할 필요가 없다. 클러스터의 표현은 각 문헌들에 출현하는 모든 용어들을 나열해 처리한다.

이 방법의 단점은 클러스터를 길게 나열하는 경향이 있어 타원형의 클러스터를 형성하기에는 적합하지만, 원형이나 분리형인 클러스터를 표현하기에는 부적합하다.

이 방법에는 Van Rijisbergen 알고리즘, SLINK 알고리즘, 최소신장 트리(Minimal Spanning Tree : MST)알고리즘이 있다. 이들은 클러스터를 구성하는데 소요되는 시간으로 O(N2)을, 기억 장소로 O(N)을 필요로 한다.

(2) 완전연결

완전 링크 방법은 문헌간의 최소 유사도를 근거로 클러스터를 형성하며, 이 방법은 싱글 링크 방법과는 달리 클러스터 내의 가장 유사도가 낮은 문헌들을 기준으로 하여 새로운 클러스터를 구성한다. 이 방법에 의해 구성된 클러스터는 모든 개체가 가장 작은 유사성으로도 서로 연결되기 때문에 완전연결이라 부른다. 작고 단단하게 묶인 클러스터가 이 방법의 특징이다.

이 방법에서도 클러스터를 표현할 때 각 문헌들에 출현하는 용어들을 나열해 표현하므로 센트로이드가 필요없으며 클러스터를 형성하는데 필요한 시간은 알고리즘에 따라 O(N2)~O(N3)시간이, 기억 장소는 O(N)~O(N2)을 필요로 한다. 이 알고리즘은 정확한 계층을 생성하기 어려운 단점과 시간적인 측면과 공간적인 측면에서 상당한 무리가 따르므로 방대한 자료 집합에 적용하기는 어렵다. 완전연결 방법에 해당되는 알고리즘은 Defays의 CLINK 알고리즘이 있다.

(3) 그룹평균 방법

그룹 평균 방법은 클러스터 대상 문헌 전체의 유사도의 평균 값을 근거로 클러스터를 형성하는 방법이다. 모든 객체들은 클러스터 간 유사성에 기여하기 때문에 느슨하게 묶인 단일연결 클러스터와 견고하게 묶인 완전연결 클러스터 사이의 중간적 구조를 나타낸다. 따라서 이 방법은 앞의 다른 방법에서 요구하는 O(N2)의 시간과 O(N)의 기억장소는 적용되지 않는다.

이 방법에 해당되는 알고리즘 중 Voorhees 알고리즘이 O(N2)의 시간과 O(N)의 기억장소에 대한 요구사항을 만족시킨다. 그 이유는 클러스터를 형성하는 센트로이드와 특정 문헌 간의 유사성이 전체 그룹평균값과 일치하기 때문이다. 이와 같은 이유에서 센트로이드는 모든 문헌 벡터들의 평균이므로, 중심 값은 O(N)의 기억장소만을 요구하며, 이 센트로이드가 클러스터들 간의 유사성을 계산하는데 이용된다. 이 방법에서도 클러스터를 표현하는데 센트로이드를 이

용하지 않고 출현한 모든 용어들을 나열해 표현하는 방법을 택한다.

(4) 와드방법

와드 방법은 최소분산방법으로 알려져 있는데. 그 이유는 각 단계에서 클러스터 쌍을 결합할 때, 문헌간의 거리를 유클리디안(euclidean)거리를 사용하여 최소 값을 갖는 것만을 연결하는 방법을 택하기 때문이다. 따라서 이 방법의 수학적 특성은 RNN(상호 밀착 이웃 : reciprocal nearest neighbor)알고리즘의 적용이 가능하다. 이 방법은 어떤 클러스터나 밀착 이웃(NN : nearest neighbor)이 존재하므로 소수의 객체쌍으로 구성된 RNN을 구성할 수 있다.

이 방법은 대칭적 계층과 동질 클러스터를 만드는 경향이 있고, 클러스터의 무게 중심에 대한 정의는 클러스터를 표현하는 유용한 방법을 제공한다. 이 방법은 클러스터 구조를 회복하는 데에는 좋으나 분리된 클러스터에 민감하고 늘어난 클러스터를 회복하는 데는 부적합하다. RNN알고리즘 역시 O(N2)의 시간과 O(N)의 기억장소에 대한 요구사항을 만족시킨다.

(5) 중앙값 방법, 중심값방법

중앙값(Median) 방법에서 클러스터는 집단 중앙의 좌표에 의해 표현되며, 가장 유사한 평균 중앙의 클러스터 쌍이 각 단계에서 새로운 클러스터를 형성하는 방법을 취한다. 중심값(Centroid) 방법 역시 동일한 방법으로 새로운 클러스터를 형성한다. 이 두 방법의 차이는 클러스터를 형성하기 위한 유사치 값을 계산할 때 클러스터의 크기(클러스터에 포함된 용어의 수)에 비례해 가중시키느냐의 여부이다. 이 방법들의 단점은 새롭게 생성된 클러스터가 클러스터 계층을 반전시키는 결과를 초래할 수 있다는 점이다.

이에 해당하는 알고리즘이 Yu(1974)가 개발한 질문식을 이용한 클러스터링 알고리즘이다. 이 알고리즘은 앞의 단일링크 알고리즘이 클러스터를 대표하는 센트로이드를 표현할 때, 포함하는 모든 색인어를 대상으로 하기 때문에 파일 구성에 많은 용량과 시간이 소요되는 단점을 개선하기 위한 것이다. 이 알고리즘의 구축방법은 싱글링크 알고리즘과 거의 유사하나 센트로이드의 대상을 클러스터에 속한 문헌들에 공통적으로 출현하는 색인어만을 대상으로 한다는 점이다. 유의 알고리즘은 "이용자들이 일정 주제의 정보를 요구할 때 그 주제에서 출현빈도가 높은 용어를 이용해 정보를 검색할 것이다." 란 가설에서 출발한다.[29]

이 알고리즘을 이용한 클러스터 구성방법은 다음과 같다.[30]

① 문헌-용어 행렬을 이용해 문헌-문헌 비유사행렬을 구성한다.

② 형성된 행렬에 기준치 'T'를 부여해 비유사계수가 기준치보다 크면 '0' [Dis(Di, Dj) > T

29) Gerald Salton, *Dynamic Information and Library Processing*, New-Jersey : Prentice-Hall, 1975. pp. 353-357.
30) T. Yu Clement, "A Clustering Algorith Based on User Queries", *JASIS*, Vol. 25, No. 4(1974), pp. 218-226

→ '0'], 그 반대의 경우에는 '1' [Dis(Di, Dj) ≦ T → '1'] 로 표시해 문헌-문헌 비유사계수 행렬을 이진 기호(binary code)로 변환시킨다.

③ 문헌간의 관계가 '1'로 되는 문헌들을 연결해 클러스터를 구성한다.

④ 이때 클러스터를 구성하는 문헌들의 공통용어들을 그 클러스터의 센트로이드로 한다.

⑤ 형성된 클러스터의 센트로이드와 기타 문헌들의 관계를 조사해 새로운 클러스터를 구성한다.

⑥ 새로 구성된 클러스터의 센트로이드를 공통 출현한 용어들로 표현해 준다.

⑦ 모든 클러스터들이 하나의 정점(頂點) 혹은 하나의 클러스터로 구성될 때까지 이 과정을 반복한다.

⑧ 형성된 모든 클러스터들을 비유사치가 낮은 수준(level)에서 높은 수준으로 차례대로 배치시켜 모든 문헌이 하나의 클러스터에 소속되도록 한다.

Yu의 알고리즘은 센트로이드의 표현을 질문식을 기준으로 표현하기 때문에 파일의 크기가 작아 비용적인 측면과 시간적인 측면에서 효과적이다. 또한 형성된 클러스터의 형태도 문헌들의 입력순서에 영향을 받지 않으며, 클러스터들의 중복이 허용된다는 장점을 가진다.

5) 알고리즘 선택기준

본 논문은 시소러스 브라우저를 설계하고자 하는 것이다. 이같은 본 논문의 목적에 해당하는 시소러스는 의미에 대해 계층적 구조를 가지므로 시소러스 브라우저의 설계에 필요한 클러스터링 알고리즘은 분할 클러스터링 알고리즘보다는 클러스터들의 계층 구축이 가능한 계층적 알고리즘이 더 적절하다. 또한 계층적 알고리즘 중 분열방법보다는 응집 방법이 효과적이므로 클러스터링 알고리즘의 평가는 계층적 알고리즘 중 응집방법에 해당되는 것으로 제한한다.

일반적으로 정보검색시스템의 효율성을 높이기 위해서 수많은 클러스터링 알고리즘 중 성능이 우수한 알고리즘을 선택하는 평가기준은 다음과 같다.[31][32][33]

① 데이터의 종류, 양, 그리고 입력순서에 관계없이 분류결과는 동일해야 한다.

② 데이터의 기술 과정에 약간의 착오가 있더라도 분류결과에 큰 영향을 초래하지 않는 안정성이 있어야 한다.

③ 클러스터를 구성하는데 소요되는 시간이 짧아야 하며, 축적 용량이 작아야 한다.

---

31) Fazli Can and Esen A. Ozkarahan, "Two Partitioning Type Clustering Algorithms", *JASIS*, Vol.35 No.3(1984. 9), p.269

32) Gerald Salton. *Dynamic information and Library Processing*. New-Jersey : Prentice-Hall, 1975. p. 329.

33) C. J. van Rijisbergen. *The Hyper-Textbook of the C.J. Van Rijsgergen's textbook on Information Retrieval*. <http : //www.dei.unipd.it/~melo/bible/documents>.

④ 이용자의 요구정보를 신속히 검색할 수 있어야 하며, 비적합한 문헌을 가능한 배제하고 많은 수의 적합문헌만을 검색해낼 수 있어야 한다.

이상과 같은 알고리즘 선택기준을 근거로 계층적 알고리즘 중 응집적 방법을 비교하면, 시간적인 측면과 저장용량 측면에서는 큰 차이가 없으며, 중심값(센트로이드) 방법에 해당하는 유의 알고리즘만이 n^2+mn번의 반복비교가 필요해 더 많은 시간과 용량이 필요한 것으로 보여진다. 그러나 유의 알고리즘은 클러스터의 표현을 문헌 쌍간의 공통 용어에 의해서만 표현하므로 용어 수가 적어 가장 신속하게 클러스터를 형성한다고 판단된다.

검색 성능이란 측면에서 평가하면, Voorhees(1986)는 1만건 이상의 대규모 문헌을 대상으로 단일연결, 완전연결, 집단평균 방법의 성능을 비교하였는데, 소규모의 데이터베이스에서는 완전연결과 집단평균 방법이 우수하였고, 대규모인 경우에는 완전연결이 우수한 결과를 나타냄을 발견하였다.[34]

검색 신속성이란 측면에서는 유의 알고리즘이 가장 우수하다. 그 이유는 다른 알고리즘은 클러스터의 표현을 소속된 모든 문헌들의 색인어들로 표현하고 있으나, 유의 알고리즘은 공통 색인어만으로 표현하고 있어, 비교대상이 적어져 신속한 검색이 가능하다.

특히 시소러스는 상하위개념과 관련어를 추출해야 한다는 관점에서 계층별로 센트로이드에 해당하는 용어들이 차례대로 누락될 수 있어 이를 근거로 상하위개념과 관련어를 자동으로 추출할 수 있는 Yu의 알고리즘이 가장 타당하다고 판단된다.

이상과 같은 성능 평가를 근거로 가장 우수하다고 판단되는 클러스터링 알고리즘은 동일 기준치에서 클러스터 간의 중복성이 무시되지만 검색이 신속하며, 파일구성에 필요한 용량과 시간적인 측면에서 우수하며, 입력문헌의 순서에 영향을 받지않는 동일한 분류결과를 제공하는 Yu의 알고리즘이 가장 우수하다고 판단된다.

3. 자동검색방법

1) 일반적 검색과정

정보검색의 효율성을 보장하기 위한 정보탐색과정의 핵심은 질의어 선정을 위한 사전탐색, 선정된 질의어들과 부울린 로직의 조합을 이용한 검색전략(검색식) 구축, 검색된 결과에 대한 평가를 근거한 피드백 탐색(사후 탐색)이다. 그러나 최종 이용자들은 특정 데이터베이스나 시스템에서 사용되는 시소러스나 주제명표목, 용어사전화일, 시스템 언어 등에 대한 이해 부족

34) E. M. Voorhees, "Implementing Agglomerative Hierachic clustering Algorithms for Use in Document Retrieval", *Information Processing & Management*, Vol. 22(1986), pp. 465-476.

은 물론, 부울린 로직을 이용한 검색결과의 축소 및 확대에 필요한 탐색전략(strategy)과 기법(tactic)등에 대한 이해가 부족하다.

따라서 이와 같은 문제점들을 해결하기 위하여 각 과정에 대한 개별적 연구와 종합적 연구들이 이루어지고 있다. 그 방법들의 결론은 연산자를 사용하지 않은 자연어 문장형태(또는 탐색 용어 형태)로 정보를 요구하고, 이 요구 정보를 파싱과 스테밍을 통해 불용어를 제외한 개념어를 추출해 초기질의벡터를 생성토록 하고, 이 벡터를 유사도 공식을 적용시켜 문헌내 출현용어를 근거한 문헌벡터와 비교해, 일정 수준 이상의 문헌만 순위를 부여해 검색하는 과정을 거치는 비부울린 탐색방법이다. 다음 장에서는 각 단계에 관련된 알고리즘을 기술하기로 한다.

(1) 질의어 확장

질문을 구성하는 방법은 시스템을 실제로 이용하는 이용자의 정보요구를 질문으로 구성하는 방법과 실험을 위하여 인공적으로 만든 질문이 있다.[35] 이 중 이용자의 정보요구를 탐색식으로 작성하여 검색하는 방법이 적합성 판정을 정확하게 할 수 있으므로 신뢰성있는 검색효율을 얻을 수 있다. 그러나 이용자의 초기 질문은 주제와 관련있는 소수의 용어로만 구성되기 때문에 효과적인 검색을 하기 위해서는 초기 탐색어 집합에 이형동의어, 동의어, 관련어 등을 추가할 필요가 있다.

이러한 과정을 질문확장 또는 질의어 확장(query expansion)이라고 하며, 그 방법은 시소러스나 의미네트웍 등과 같은 정보원을 사용해 확장되는 지식기반 확장 방법[36]과 초기질의벡터의 탐색으로 검색된 문헌 중 적합문헌에 출현한 용어들을 사용해 확장되는 탐색결과 기반 확장 방법 등이 있다. 탐색결과 기반확장은 피드백 탐색 알고리즘에서 논하기로 한다.

그러나 시소러스를 이용한 대부분의 현존 시스템들은 시소러스를 통해 제한된 질의 확장과 상호작용적 접근만을 해결하고 있다. 이는 시소러스가 갖고 있는 계층(hierachical), 연관(associative), 등가(euivalence)관계를 통한 용어간의 의미적 네트웍 관계를 간과했기 때문이다.

이에 대한 해결 방법은 의미론적 색인 공간(semantic index space)에서 용어들간의 의미적 밀접성(semantic closeness)의 측정이다.[37] 의미적 밀접성 안에는 용어 간의 거리란 유전적 개념이 내포되어 있다.

---

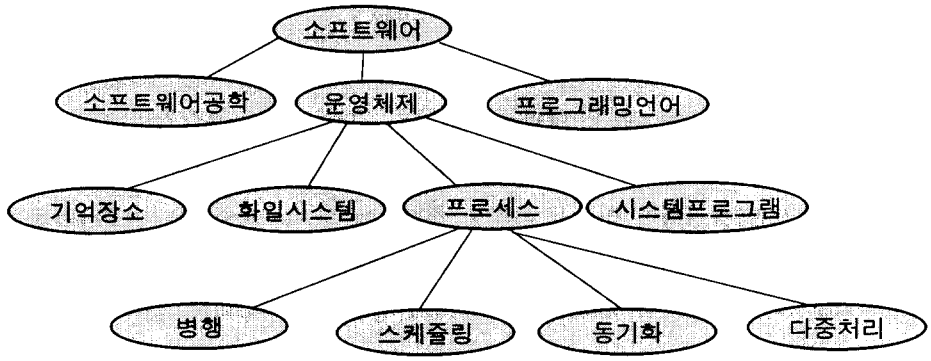
35) J. Taegu, "The Pragmatics of Information Retrieval Experimentation Revisited", *IPM*, vol. 28, no. 4(1992), p. 476.

36) Helen J. Peat & Peter Willett, "The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems", *JASIS*, Vol. 42, No. 5(1991), p. 378.

37) D. Cunliffe, C. Taylor, D. Tudhope, "Query-based navigation in semantically indexed hypermedia", 《Proceedings of the ACM Conference on Hypertext》 Vol. 8(1997), pp 87-95.

질의어를 확장하는 과정은 시소러스와 대조하여 일치하는 명사가 있으면 이 명사와 협의관계에 있는 명사들을 ORing시켜 확장시키는 방법을 택하고 있다. 이때 일치되는 바로 밑의 용어들만 확장시키고 그 이하의 용어들은 확장시키지 않는 이유는 모든 협의어들이 전부 확장되는 경우에는 확장되는 용어의 수가 너무 많고 이에 따라 너무 특정한 용어에까지 확장되기 때문이다. 또한 만약에 시소러스에 일치하는 명사가 시소러스의 단말 노드라면 확장할 협의어가 없으므로 그 용어의 형제 노드들을 ORing 하여 질의어를 확장한다.

질의어 확장 방법과 과정을 계층화된 용어개념 <그림 3>을 근거로 설명하면 다음과 같다.



〈그림 3〉 시소러스를 이용한 질의어 확장

그림의 예에서 질의어가 (운영체제 and 동기화)라면 시소러스에서 질의어의 용어와 일치하는 용어를 찾는다. '운영체제'의 바로 밑의 협의어들은 '기억장소', '화일시스템', '프로세스', '시스템프로그램'이므로 이들을 ORing한다. 다음 질의어 '동기화'는 바로 밑의 협의어가 없으므로 형제노드인 '병행', '스케쥴링', '다중처리' 등을 ORing 한다. 그 결과 질의어 확장은 ((운영체제 or 기억장소 or 파일시스템 or 프로세스 or 시스템프로그램) and (동기화 or 병행 or 스케쥴링 or 다중처리))와 같이 된다.

(2) 검색 알고리즘

최종 이용자가 겪는 탐색의 어려움은 사용하는 특정 데이터베이스나 시스템에서 사용되는 시소러스나 주제명표목, 용어사전화일, 시스템 언어 등에 대한 이해 부족은 물론, 부울린 로직을 이용한 검색결과의 축소 및 확대에 필요한 탐색전략(strategy)과 기법(tactic)등에 대한 이해가 부족한데서 기인한다. 특히 최종 이용자들이 정보요구를 부울린 로직을 이용한 탐색식으로 표현하는데 겪는 어려움과 부울린 로직을 이용한 검색방법의 단점을 보완하기 위해서 부울린

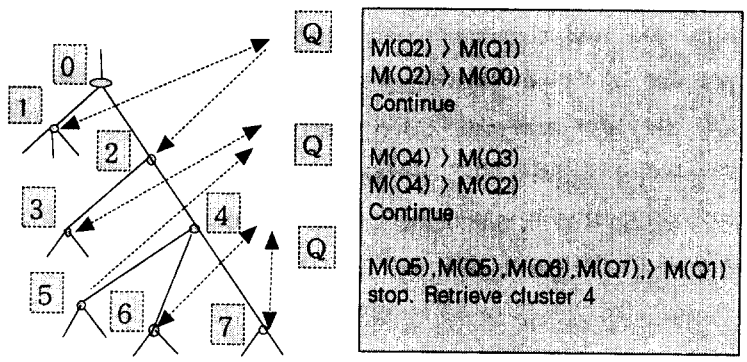
로직을 사용하지 않는 검색방법이 연구되고 있다.

검색방법의 종류는 크게 부울린 탐색(Boolean search)과 매칭 함수(Matching functions)에 의한 탐색 방법으로 나뉘어진다.[38] 부울린 탐색은 and, or, not 등의 연산자를 근거로 정보를 검색하는 방법이며, 매칭 함수(matching fuctions)를 이용한 탐색은 질의를 문헌이나 클러스터와의 연관성을 근거로, 즉 Dice 계수, cosine 계수나 Tanimoto 계수 등의 연관성 측정법(association measure)을 적용해 일정 기준치를 통과하는 문헌만을 원하는 정보라고 판단하여 검색하는 방법이다. 매칭함수에 의한 탐색은 순차탐색(serial search), 클러스터탐색(clustered based search)이 존재한다.

순차탐색은 모든 문헌의 용어들과 질의 용어를 매칭함수로 비교해 정보를 검색하는 것으로 검색결과는 특정 기준치(threshold value)에 의해 제공되거나, 매칭함수에 의한 문헌의 서열을 근거한 절단서열위치(cutoff rank position)에 의해 제공된다. 클러스터 탐색은 클러스터 알고리즘에 의해 구성된 클러스터의 표현(representatives)을 매칭함수로 비교해 정보를 검색한다. 그 방법은 하향식(top-down) 탐색법과 상향식(bottom-up) 방법이 있다.

하향식 방법은 최상위 레벨 클러스터에서 시작하여 하위레벨 순으로 매칭함수 값을 비교하는 방법으로서, 검색결과의 제시를 위해 중단 규칙(stopping rule)이 필요하다. 중단규칙은 클러스터와 질의를 비교한 매칭함수가 현 계층 클러스터의 매칭함수가 바로 위 계층의 매칭함수보다 작아질 때 중단토록 하는 방법을 택하고 있다.

검색결과는 바로 위 계층의 클러스터에 소속된 모든 문헌들을 제공된다. 이 방법의 단점은 각 node에 의해 기하급수적으로 증가하는 모든 클러스터를 비교해야 하므로 시간이 많이 걸린다는 점이다. 그 과정은 다음의 그림과 같다.



〈그림 4〉 매칭함수를 적용한 탐색 트리내에서 중단규칙을 적용한 검색방법

38) C. J. van Rijisbergen. *The Hyper-Textbook of the C.J. Van Rijsgergen's textbook on Information Retrieval.* <http : //www.dei.unipd.it/~melo/bible/documents/98.html>.

상향식 방법은 최하위 레벨 클러스터에서 시작하여 상위레벨 순으로 매칭함수 값을 비교하는 방법으로서, 검색결과의 제시를 위한 매칭함수의 비교는 중단 규칙을 사용하는 대신, 임의의 최대문헌 수를 초과하지 않으면서 초기 노드(starting node)를 포함하는 최대 형태의 클러스터를 발견할 때까지 계속한다. 발견된 클러스터에 속한 문헌들이 검색결과로써 제시된다. 이 방법의 문제점은 최적의 초기 노드를 찾는 방법의 어려움이다.

(3) 피드백 탐색 알고리즘

매칭함수에 의한 클러스터 탐색에 의해 초기 질의어에 대한 탐색이 종료되었을 때, 검색결과가 만족스럽지 못할 경우 새롭게 탐색을 수행해야 한다. 이를 피드백(feedback) 탐색, 또는 탐색결과에 근거한 질의확장(query expansion search based on search results)이라고 한다. 검색결과가 만족스럽지 못한 이유는 주로 이용자들이 요구를 정확히 제시하지 못했기 때문에 발생한다. 일반적으로 이용자는 시행착오의 과정을 거쳐 정보요구를 구체화하므로 탐색결과를 근거로 정보요구의 재형성 과정이 필요하다. 재형성시 참고할만한 정보는 데이베이스내에서 탐색어의 출현 빈도, 검색된 문헌의 수량, 대체 가능한 용어와 관련 어휘, 적합하다고 인정되는 서지와 서지내의 색인어휘, 계층화된 어휘 사전 등이다.

탐색결과를 근거로 자동으로 질의를 확장하여 정보를 탐색하는 방법은 매칭함수를 이용해 구성이 가능하다. 자동 피드백 탐색 방법은 먼저 초기 질의어에 의해 검색된 문헌들에 대해 매칭함수를 이용한 적합성 서열화가 이루어져야 하며, 이를 근거로 가장 적합하다고 판단되는 10% 이내의 문헌에 출현하는 용어를 근거로 초기질의를 확장 또는 수정하여 탐색을 수정하는 과정을 거친다. 적합성피드백에 의해 초기 질의를 수정 또는 확장하는 방법은 다음과 같은 방법이 있다.[39]

첫째, 질의어 자동 수정 방법(Automatic Query Modification) : 초기 질의어에 대한 확장은 하지 않고, 초기질의어의 각 용어에 대한 적합성 가중치만 재산출하여 탐색을 수행하는 방법.

둘째, 질의어 자동 확장 방법(Automatic Query Extension - All) : 초기 질의어에 검색된 문헌 중에서 상위 적합문헌에 출현하는 모든 용어를 이용하여 탐색을 수행하는 방법.

셋째, 질의어 자동 선별 확장 방법(Automatic Query Extension - Sellect) : 초기 질의어에 검색된 문헌 중에서 상위 적합문헌에 출현하는 모든 용어를 가중치순으로 순위를 부여해 상위 n개의 용어만을 이용하여 탐색을 수행하는 방법.

넷째, 탐색자 개입 질의어 확장 방법(Interactive Query Expansion) : 초기 질의어에 검색된 문헌 중에서 상위 적합문헌에 출현하는 모든 용어를 가중치순으로 순위를 부여하고, 이를 이

39) 노정순, "탐색결과에 근거한 자연어질의 자동확장 및 응용에 관한 연구 고찰", 《정보관리학회지》, Vol. 16, No. 2(1999), pp. 69-71.

용자가 선택한 질의어만으로 새롭게 탐색을 수행하는 방법.

이들 방법들의 성능 평가에 대한 선행 연구의 결과는 질의어 수정보다는 확장방법이 우수하며, 확장 방법 중에서는 탐색자 개입 질의어 확장 방법이 가장 우수하다고 보고하고 있다.

2) 클러스터링 시소러스 브라우저를 이용한 검색방법

본 장에서의 검색방법은 앞의 클러스터링에 의해 구축된 시소러스를 이용해 질의어를 확장하는 방법을 논의할 것이다. 또한 개념이 확장되어 제시된 질의어들 중 이용자가 선택한 질의어를 기준치를 이용해 검색식을 자동으로 확장하거나 축소하는 방법을 통해 이용자들의 직접적인 사고과정이 없이도 자동으로 부울린 로직을 이용한 탐색식을 구성하여 탐색을 수행하는 방법에 대해 논의할 것이다. 또한 검색결과를 제시할 때도 이를 근거한 적합도 순위를 부여하는 방법에 대해서도 논의될 것이다.

(1) 질의어 확장

컴퓨터에 의한 정보검색의 목적은 비적합한 문헌을 가능한 배제하고 많은 수의 적합한 문헌들을 검색해내는 것이다. 이를 위해서는 문헌정보를 표현할 때, 그 표현방법을 이용자의 요구에 적합하도록 표현해야 한다. 과거 수작업 검색에서의 색인 작업도 문헌에 할당되는 용어의 선정을 이용자가 질의를 구성할 때 동일한 용어를 사용할 것이란 것을 전제로 구성되어 왔다. 따라서 컴퓨터에 의한 자동검색도 자동색인에 의해 형성된 색인어를 위주로 사용해야 검색 성능이 우수해질 것이다.

클러스터링을 이용해 구축된 시소러스 브라우저에서의 질의어 확장 방법은 문헌에서 실제로 빈번히 사용되고 있는 용어로 질의어를 확장하는 것이다. 앞의 클러스터링 과정에서 생성된 용어군들은 노드 값에 따라 분리되는데, 이 노드들은 부모 노드와 자식 노드의 관계가 형성된다. 즉 상위 노드의 센트로이드 값은 하위 노드의 센트로이드 값보다 상위개념이 되며, 하위 노드의 센트로이드 값은 하위개념이 된다. 이는 시소러스의 광위어와 협의어의 관계로 설명된다.

질의어를 확장하는 과정은 질의 내의 출현 명사를 계층화된 시소러스 브라우저의 노드 값에 해당하는 클러스터 표기(representative, centroid)에 포함되는 용어들과 대조하는 것으로 시작한다. 시소러스 브라우저는 부울린 로직의 and 기능을 통해 입력된 초기 질의어들이 전부 수록되어 있는 센트로이드들을 매칭함수에 의해 서열화해 제시한다. 물론 이 과정에서 이용자의 간섭없이 자동으로 질의어를 확장할 수 있으나, 검색의 결과는 이용자가 스스로 선택한 질의어에 의해 더 만족될 수 있기 때문에 관련 질의어들을 서열화해 제시토록 한다.

정보요구에 대한 질의어 확장 중 초기 질의어 확장에 관심을 갖는 것은 지식기반데이터베

이스(시소러스 브라우저)를 이용한 정보검색에서는 문헌탐색과 관련된 용어들이 자동화된 정보검색시스템의 구축과정에서 사전에(정보탐색 수행 이전에) 매칭기법과 클러스터링 기법에 의해 개념의 매핑이 이루어졌기 때문이다. 또한 센트로이드에 포함된 용어들은 용어간의 계층관계, 동위관계, 등가관계 등의 개념간의 매핑이 조화를 이룬 분류기호와 같은 성격의 용어들이기 때문이다.

(2) 검색 알고리즘

클러스터링을 이용해 구축된 시소러스 브라우저에서의 검색방법은 최종 이용자가 겪는 탐색의 어려움 중에서 부울린 로직을 이용한 검색결과의 축소 및 확대에 필요한 탐색전략(strategy)과 기법(tactic) 등에서 발생하는 어려움을 해결한다. 이 과정에서 적용되는 알고리즘은 부울린 로직, 매칭함수, 기준치를 근거한 탐색방법이다.

매칭함수는 질의어 확장에 적용되어 시소러스 브라우저의 개념 노드에 해당하는 센트로이드가 포함하는 용어들을 제시하기 위해 사용된다. 매칭함수는 초기 질의어에 수록된 용어들의 일부가 센트로이드에 포함되어 있지 않더라도 가장 유사한 센트로이드를 검색하도록 해주며, 검색된 확장 질의어 대상들의 서열을 제시하기 위해 적용된다.

부울린 로직은 확장된 질의어를 실제 문헌내에 수록되어 있는 용어들과 대조해 검색하는 과정에서 주로 적용된다. 이 과정에서 질의어들은 모두 'and'로 결합해 탐색을 수행한다. 검색된 결과들의 서열화는 최신성을 근거로 제공된다.

기준치를 근거한 탐색방법은 피드백 탐색방법에 해당하므로 다음 장에서 논하기로 한다.

(3) 피드백 탐색

피드백 탐색은 앞의 검색방법을 적용한 결과가 너무 광범위한 내용이거나 협소한 내용이기 때문에 이용자가 만족하지 못할 때 적용한다. 그 방법은 앞장에서 설명한 초기탐색 결과 중 적합 문헌내의 출현용어들을 근거로 사후탐색을 하는 피드백 알고리즘과는 차이가 난다. 클러스터링을 이용한 시소러스 브라우저에서의 피드백 탐색은 다음과 같다.

첫 번째 방법은 기준치(threshold value)를 적용하는 방법이다. 이 방법은 기준치에 의한 매칭함수를 적용시켜 질문식내의 용어들을 자동으로 순열화해 새로운 질의어 집합들을 구성하며, 각 질의어 집합내의 용어들은 'and'로, 집합끼리는 'or'로 연결된 탐색식을 자동으로 구축해 검색을 수행한다. 이 방법은 주로 재현율을 확장하기 위해 적용될 것이다.

두 번째 방법은 인접 센트로이드를 제시하는 방법으로서, 즉 탐색에 적용한 확장 질문식보다 하위 질문식과 상위 질문식을 제시해 이용자가 다시 질문식을 선택하도록 유도하는 방법이다. 이는 재현율과 정도율을 선택해 적용될 것이다.

세 번째 방법은 브라우저에 수록되어 있지 않은 특정 용어를 초기 질문식과 'and'로 연결시키는 방법으로서 이용자의 세부적 정보 욕구, 즉 정도율을 강화시키기 위한 방법이다.

이상과 같은 클러스터링을 이용한 시소러스 브라우저에서의 피드백 탐색은 이용자의 만족도를 향상시켜 줄 것이다. 그러나 저자의 판단에 의하면, 초기의 질의어 확장시 이용자가 정확한 질의어를 선택했다면 대부분의 이용자는 만족할만한 수준의 검색결과를 획득할 수 있을 것이다.

Ⅳ. 결 론

최종 이용자는 검색전략과 검색기법의 미숙에 의해 검색결과에 만족하지 못하는 경우가 많다. 이에 대한 해결책은 질의어 선정을 위한 사전탐색, 질의어와 관련이 높은 정보들이 수록되어 있는 데이터베이스의 선택, 선정된 질의어들과 부울린 로직의 조합을 이용한 검색전략(검색식) 구축, 검색식에 근거한 온라인 탐색, 검색된 결과에 대한 평가를 근거한 피드백 탐색(사후 탐색) 등의 정보탐색 과정에서 발생하는 문제들을 사전에 해결하는 방법이어야 한다.

본 논문의 가설은 이용자의 만족도를 극대화시킬 수 있는 검색시스템은 저자(연구자)들이 전문에서 사용하고 있는 용어와 탐색자의 질의용어를 일치시킬 수 있는 방법으로 구축된 검색시스템이어야 한다는 것이다. 따라서 본 연구의 논제가 '클러스터링을 이용한 시소러스 브라우저에 대한 이론적 모형'임에도 불구하고 기술된 내용은 내용 분석, 정보구조, 질의 형성, 질의 평가 등 자동정보검색의 전 분야를 망라하고 있다. 그 이유는 용어의 동시 출현빈도를 이용한 자동정보검색 방법이 가설을 해결할 수 있을 것이라 판단하였기 때문이다.

각 장에서 분석한 내용의 결과는 다음과 같다.

첫째, 시소러스 브라우저는 전통적인 수작업 과정에 의해 만들어진 시소러스를 단순히 컴퓨터에 이식시킨 수준에서 벗어나 저작물 내의 출현 가능성이 높은 용어들을 이용해 계층화가 이루어진 형태로 제공되어져야 한다.

둘째, 시소러스 브라우저의 기능과 형태는 사전탐색, 데이터베이스 선택, 탐색전략 구축, 온라인 탐색, 사후탐색 등 정보검색과정의 전단계에 걸쳐 중요한 영향을 미치는 핵심적 시스템으로서의 기능을 갖추어야 한다. 따라서 시소러스는 기존의 기능 뿐 아니라 온라인 환경에서의 정보검색의 효율성 제고를 위하여, 시스러스의 지식베이스로서의 가능성을 구현한 인간중재 전문가 시스템의 기능을 갖춰어야 하며, 그 형태는 이용자가 편리하게 사용할 수 있도록

이용자 지향적 시스템이어야 한다.

셋째, 위와 같은 기능을 갖추기 위해서 시소러스 브라우저에 대한 연구는 내용 분석, 정보구조, 질의 형성, 질의 평가 등에 깊은 연관성을 갖고 있음을 인식하고 자동정보검색의 전 분야에 대한 심도깊은 연구를 해야 할 것이다.

넷째, 색인은 질의어 구성시 문헌내에서 저자가 사용하는 용어와의 일관성을 보장하는 문헌보증(literary warrant)의 역할을 해야하므로 철저성과 망라성을 겸비해야 한다. 따라서 통계적 기법, 언어학적 기법, 문헌구조적 기법이 조합되어 의미있는 용어를 추출하는 과정인 자동색인은 시소러스 브라우저 구축에서 가장 중요한 요소이다. 그러나 한글에 대한 자동색인 추출과정시 문자열 식별방법, 불용어목록, 스테밍 과정, 복합명사 처리 등의 기준에 대한 불명확성과 비공개성 때문에 시소러스 브라우저의 초기단계에서 많은 불편이 따른다. 따라서 색인어의 철저성과 망라성을 보장할 수 있는 한글에 대한 자동색인 알고리즘의 표준화가 시급하다.

다섯째, 자동색인을 검증하기 위해서는 원문이 있어야 하는데, 그 원문은 전문 전체를 요약한 초록이 주 대상이 될 것이다. 그러나 초록의 내용이 한글보다는 영문으로 작성한 경우가 많아 특정 주제에 대한 자동색인은 물론, 시소러스 구축의 어려움으로 존재한다. 또한 일부 데이터베이스의 경우 영문초록을 능력이 부족한 번역자를 통해 한글로 직역한 경우가 많기 때문에 특정주제에 대한 내용으로 미흡한 경우가 많다. 따라서 모든 학술지와 학위 논문 등의 연구 업적물에 저자들의 한글초록과 영문초록을 반드시 병행해 기재해야 할 것이다.

여섯째, 시소러스 브라우저를 구축하기 위한 용어 클러스터링 알고리즘은 반드시 용어들의 계층화된 표현이 가능하도록 계층 클러스터링 알고리즘을 적용해야 하며, 클러스터를 형성토록 한 용어의 표현은 클러스터의 중심값인 센트로이드에 해당되는 용어로 표현함에 의해 저장공간과 검색의 속도를 보장할 수 있을 것이다.

일곱 번째, 검색방법은 종래의 부울린로직을 적용한 방법보다 매칭함수를 응용한 검색방법이 더 유용할 것이며, 이를 통해 질의어의 확장, 검색결과의 서열화, 피드백 탐색등의 검색과정을 이용자지향적 시스템으로 제공할 수 있어 이용자의 만족도를 극대화할 수 있을 것이다.

여덟 번째, 시소러스 브라우저는 최신 용어들의 출현과 특정 용어들의 중요도 비중의 변화에 의해 검색결과가 차이가 발생할 수 있으므로 주기적으로 그 내용을 갱신해야 하며, 이같은 갱신에는 종래의 구축방법보다 클러스터링을 이용한 시소러스 브라우저의 갱신방법이 더 효과적일 것이다.

아홉번째, 시소러스 브라우저 형성과정 중 형성된 용어들간의 관계는 의미망 구축이 가능해 지식기반 데이터베이스로서의 기능을 발휘할 수 있을 것이다.

이상의 분석결과를 근거로 '클러스터링을 이용한 시소러스 브라우저' 구축의 이상적 모형은 다음과 같다.

첫째, 시소러스에 수록된 용어는 특정 주제 문헌에 출현한 용어만으로 구성한다.

둘째, 문헌 내에서 자동색인을 추출하는 과정은 색인어의 철저성과 망라성을 보장할 수 있는 알고리즘으로 적용되어야 한다.

셋째, 클러스터링 알고리즘은 용어간의 계층화, 동의관계, 등가관계를 추출할 수 있는 계층 알고리즘으로 하며, 검색의 신속성과 저장성을 고려해 중심값(센트로이드) 알고리즘을 적용한다.

넷째, 검색방법은 초기의 질의어 확장을 핵심으로 하고, 확장 질의어 검색방법은 인터페이스 기능을 이용한 매칭함수를 통해 관련된 대상 질의들을 제시하고, 이용자가 직접 탐색을 수행할 질의어들을 선택하는 방법을 택한다. 검색식의 수행은 질의내에 포함된 모든 용어들을 and로 결합한 탐색식으로 수행하며, 결과의 제시는 최신성을 고려해 순서화해 제시한다.

다섯째, 피드백탐색은 재현율을 확장하기 위한 기준치(threshold value) 적용 방법, 정도율을 강화시키기 위한 방법으로 브라우저에 수록되어 있지 않은 특정 용어를 초기 질문식과 'and'로 연결시키는 방법, 재현율과 정도율을 선택적으로 수용할 수 있는 인접 센트로이드를 제시하는 방법을 통해 해결한다.

이상과 같은 본 논문의 결과는 추후 발표될 '클러스터링을 이용한 시소러스 브라우저의 구축'에 실제적으로 적용되어 이론적 결론이 입증될 것이다.

참 고 문 헌

남영준. "색인어형태분석에 의한 한국어 자동색인기법 연구". 박사학위논문, 중앙대학교 대학원, 1994.

노정순. "탐색결과에 근거한 자연어질의 자동확장 및 응용에 관한 연구 고찰". 《정보관리학회지》 vol. 16, no. 2(1999), pp. 49-80.

서휘. "정보검색을 위한 인버티드화일과 클러스터화일의 비교분석". 석사학위논문, 중앙대학교 대학원, 1986.

이재윤, 김태수. 『Wordnet과 시소러스, 언어정보 연찬회 발표논문집』 제11권(1998), pp. 1-19.

정영미, 이재윤. "한국어 텍스트 내 용어연관성 분석을 위한 기초 연구". 《한국정보관리학회 학술대회 논문집》 제5권(1998), pp. 243-246.

한상길. "시소러스 용어관계의 확장에 관한 연구". 박사학위논문, 중앙대학교 대학원, 1999.

Frakes, William B. & Baeza-Yates, Ricardo. 『Information Retrieval : Data Structure & Algorithms』. New Jersey, Prentice Hall, 1992.

Jones, Susan and others. "Interactive thesaurus navigation : intelligence rules OK?." *JASIS,* vol. 46, no. 1(1995). pp. 52-59.

Lancaster, F. W. *Vocabulary Control for Information Retrieval.* 2nd ed. Arlington, Virginia, Information Resources Press, 1986.

Miyamoto, S. "Information Retrieval Based on Fuzzy Associations." *Fuzzy Sets and Systems,* vol. 38(1990), pp.191-205.

Peat, Helen J., & Willett, Peter. "The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems." *JASIS,* vol. 42, no. 5(1991), pp. 378-383.

Salton, Gerald. *Dynamic Information and Library Processing,* New-Jersey, Prentice-Hall, 1975.

Van Rijisbergen, C. J. *The Hyper-Textbook of the C.J. Van Rijsgergen's textbook on Information Retrieval* <http : //www.dei.unipd.it/~melo/bible/documents>.

Voorhees, E. M. "Implementing Agglomerative Hierachic clustering Algorithms for Use in Document Retrieval." *Information Processing & Management,* vol. 22(1986), pp. 465-476.

Weinberg, B. H. "Library classification and information retrieval thesauri : comparison and contrast." *Cataloging and Classification Quarterly,* vol. 19, no.3/4(1995), pp. 23-44.

Tudhope, Douglas. *Position Statement: Adaptive navigation tools for thesaurus-based retrieval in cultural heritage applications.* <http : //wwwis.win.tue.nl/ah98/Tudhope.html>