시소러스 브라우저 자동구현을 위한 Metadata를 이용한 색인어 처리방안에 대한 연구

A Theoretical Study on Indexing Methods using the Metadata for the Automatic Construction of a Thesaurus Browser

서 휘(Whee Seo)*

<목 차>

초 록

본 연구에서는 시소러스 브라우저를 자동으로 구성하기 위한 방법에 대한 이론적인 연구와 함께 시소러스 브라우저 구성과정의 핵심인 자동색인과 용어 간 계층을 자동으로 형성하는 클러스터링 알고리즘에 대한 선행 연구결과를 제시하였다. 그리고 웹 문헌에서 전통적인 종이 형태 문헌의 서지사항에 해당하는 메타데이터를 분석하고 이를 처리하는 방안을 조사함에 의해 웹 문헌에서 색인어를 자동으로 추출할 수 있는 방안에 대하여 연구하였다. 또한 대부분의 웹 문헌에 메타데이터가 수록되어 있지 않음에 착안하여 기존의 웹 문헌에 메타데이터 자동 편집기를 이용하여 메타데이터를 수록하는 방안에 대한 연구결과를 제시하였다.

주제어 : 시소러스, 클러스터링 알고리즘, 메타데이터, 자동색인

Abstract

This paper is intended to present the theoretical analyses on automatic indexing, which is vital in the process of constructing a thesaurus browser, and clustering algorithms to construct hierarchical relations among terms as well as the methods for the automatic construction of a thesaurus browser. The methods to select the index term automatically in the web documents are studied by surveying the methods for analyzing and processing metadata which conforms to bibliographical roles of traditional paper documents in web documents. Also, the result of the study suggests to adding or involving the metadata in web documents, using the metadata automatic editor because metadata is not listed in most of the web documents.

Key Words : thesaurus, clustering algorithm, metadata, automatic indexing

* 창원전문대학 문헌정보과 조교수(drs733m@changwon-c.ac.kr)
· 접수일 : 2004. 11. 21 · 최초심사일 : 2004. 12. 5 · 최종심사일 : 2004. 12. 10

Ⅰ. 서론

현재 인터넷에는 수많은 정보들이 유통되고 있다. 그럼에도 불구하고 아직까지 이용자들이 만족하는 정보를 제공하지 못하고 있다. 이와 같은 원인은 인터넷의 정보 탐색기법이 웹 문헌의 원문을 대상으로 하는 키워드 검색방법의 수준에 머물러 있기 때문이다. 이와 같은 문제점은 미래의 검색엔진인 시멘틱 웹(Semantic Web) 시스템이 구축된다면 해결될 수 있을 것이다.

왜냐하면 시멘틱 웹에서는 내장되어 있는 온톨로지(Ontology)를 이용해 추론까지 가능한 검색엔진을 구축할 수 있기 때문이다. 따라서 보다 효율적인 정보검색을 위해서 수많은 연구자들이 웹 기반의 온톨로지를 구축하기 위한 노력을 계속하고 있다. 이와 같은 노력의 결과로 소수의 정보검색사이트에서 온톨로지를 기반으로 한 정보검색 서비스를 시도하고 있다. 그러나 아직까지 웹 상에서 대용량의 보편적인 정보검색 서비스로 온톨로지가 활용되고 있는 사례는 찾아볼 수 없다.

따라서 본 연구에서는 미래의 검색엔진인 시멘틱 웹을 구축하는 핵심기술인 온톨로지를 자동으로 구성하기 위한 전단계로서 용어 간의 계층관계를 자동으로 구성할 수 있는 시소러스 브라우저 자동구축방법에 대한 이론적인 연구를 수행할 것이다. 또한 이와 같은 이론적 연구를 통해 웹 문헌에서 전통적인 종이 형태 문헌의 서지기술사항에 해당하는 메타데이터를 분석함을 통해 해당 웹 문헌의 핵심 주제를 나타내는 색인어를 자동으로 추출할 수 있는 방법을 제시할 것이다.

본 연구의 목적은 웹 문헌의 주제를 나타내는 메타데이타를 이용하여 시소러스 브라우저를 자동으로 구현하는 과정 중, 첫 단계인 웹 문헌에서 색인어를 자동으로 처리하는 방안에 대한 연구이므로 실험은 웹 문헌에서 색인어를 처리하는 방안으로 국한한다.

Ⅱ. 시소러스 브라우저의 필요성

1. 시소러스 브라우저의 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

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.

병행해 사용되는 것과는 달리 온라인서비스에서 최종 이용자가 검색과정에서만 사용하기 위한 목적(자연어 시스템에서 어의적으로 관련된 용어나 동의어를 통제하기 위한 목적)으로 만들어졌기 때문에 탐색용 시소러스[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 질문식 작성과 탐색주제 표현방법의 불명확성, 문헌에 부여된 색인어의 다양성, 정보검색시스템의 복잡성 등을 해결함을 통하여 최종 이용자가 만족할만한 주제문헌을 찾도록 지원해 주는 것이다.

따라서 시소러스 브라우저는 탐색자가 단지 하나의 용어를 입력하더라도 이를 근거로 용어의 의미 네트웍에 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.

2. 시소러스 브라우저의 필요성

1) 질의어 확장용 시스템

현재의 온라인정보서비스는 서지사항이나 초록 등의 2차 정보서비스 수준에서 벗어나 원정보인 전문(full-text) 전체를 수록하고 있는 전문데이터베이스 서비스가 보편화되고 있다. 이와 같은 변화 발전의 이유는 컴퓨터 관련 기술의 발전과 가격의 저하에 따라 대용량의 정보인 인쇄·영상·음성매체 정보들이 디지털로 표현이 가능하며, 이용자들의 정보요구가 신속성을 보장하기 위해 온라인으로 1차정보원(디지털 원문정보)을 이용하려는 욕구가 증가하였기 때문이다.

그러나 대부분의 전문데이터베이스가 사전에 색인작업을 수행하지 않기 때문에 탐색자에게 큰 부담으로 작용한다. 그 이유는 탐색식 구성시 용어의 조합을 적합문헌에 출현하는 정확한 용어를

---

2) W. Schmitz-Esser, "New Approaches in Thesaurus Application," *International Classification*, Vol.18, No.3(1991), pp.144-145.

3) F. W. Lancaster, 정보검색시스템, 윤구호, 김태승 공역(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.

예측해 조합해야 하지만, 특정주제에 대한 포괄적인 검색을 하기 위하여 제공되는 동의어, 계층어, 관련어를 예측하는 것은 저자의 저작유형이 다양하여 탐색자가 이러한 용어를 생각해내기는 무리이기 때문이다.[7]

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.[8]

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

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.[9] 이같은 결과에 대해 Borgman은 최종 이용자가 특정 데이터베이스나 시스템에 적용된 시소러스나 주제명표목, 용어사전화일, 시스템 언어 등에 대한 이해 부족은 물론, 불리안 로직(Boolean Logic)을 이용한 검색결과의 축소 및 확대에 필요한 탐색전략(strategy)과 기법(tactic)등에 대한 이해가 부족한데서 기인한다고 주장하고 있다.[10]

일반적으로 정보탐색의 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 초기질의어와 초기 검색결과를 이용해 계층어휘와 동의어 등의 관련 어휘를 추출하는 과정이다. 그러나 이 과정에 의해 추출된 확장된 어휘를 이용하더라도, 일반 이용자는 탐색전략 구축방법(and, or, not과 같은 불리안 로직을 적용시키는 방법)에 익숙하지 못하기 때문에 정확한 탐색식을 구축하지 못하고 있다.

---

7) F. W. Lancaster, 정보검색시스템, p.313.
8) Moid A. Siddiqui, "Full-Text Database," *Online Review*, Vol.15, No.6(1991), p.369.
9) E. J. Mckinin. et al., "The Medline/full-text Tesearch Project," *JASIS*, Vol.42, No.4(1991), p.303.
10) C. L. Borgman, "Why are Online Catalogs still Hard to Use?" *JASIS*, Vol.47, No.7(1996), pp.493-503.

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. Rowley는 시소러스의 기능 확장을 다음과 같이 설명하고 있다. "이용자 인터페이스는 GUI 형태로 제공되는 레코드나 탐색프로화일을 화면에 제시해 줌으로써 스크린에서 시소러스를 찾거나 통제어휘의 리스트를 보는 것이 가능하다. 또한 자연어 탐색에 대한 지능 인터페이스를 제공하는 지식베이스로서 시소러스를 이용한다. 이와 같이 시소러스는 단어들 사이의 관계를 정의하고, 확장하거나 축소하는 등 다양한 방법으로 이용자 탐색을 향상시키는 시스템으로 이용할 수 있다"[11]

따라서 본 연구에서 수행하는 클러스터링을 이용한 시소러스 브라우저는 불리안 로직과 비불리안 로직(매칭함수에 의한 확률 검색방법)이 결합된 검색방법이므로 최종 이용자가 탐색전략과 기법 등에 대한 이해 부족 때문에 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.

B. H. Weinberg는 시소러스가 이용자 지향적 시스템으로의 발전 가능성에 대하여 시스템 입장의 색인단계보다 이용자 입장의 탐색단계에서 확장이 가능하며, 특히 인공지능과 하이퍼텍스트 분야에서 필수적인 지식베이스의 구축을 위해서 반드시 시소러스 구축이 선행되어야 한다고 주장하고 있다.[12] 또한 Susan Jones는 계층구조와 같은 메뉴 인터페이스를 이용한 검색자와 문헌간의 연결기능과 지능형 정보검색시스템을 제공하여 초기질의어 확장 기능을 갖춘 인간중재 전문가시스템(Human Intermediary Expert System)의 구축에 시소러스의 검색기능이 적용된다고 주장하고 있다.[13] 또한 최석두는 시소러스가 자연어 처리시스템의 기본사전, 용어사전, 다국어 대역사전, 다국어 동적 시소러스로의 확장이 가능하다고 주장하고 있다.[14]

---

11) J. Rowley, "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.

12) B. H. Weinberg, "Library Classification and Information Retrieval Thesauri : Comparison and Contrast," *Cataloging and Classification Quarterly*, Vol.19, No.3/4(1995), p.39.

13) Susan Jones et al., "Interactive Thesaurus Navigation : Intelligence Rules OK?" *JASIS*, Vol.46, No.1(1995), p.52.

따라서 이용자 지향적 시스템의 성능을 갖춘 시소러스는 기존의 기능 뿐 아니라 온라인 환경에서의 정보검색의 효율성 제고를 위하여, 지식베이스로서의 가능성을 구현한 인간중재 전문가 시스템의 기능을 갖춰야 한다. 또한 시소러스는 사전탐색(Pre-searching), 데이터베이스 선택(DB Selection), 탐색전략 구축(Searching Strategy Construction), 온라인 탐색(Online Searching), 사후탐색(Post-searching) 등 정보탐색과정의 각 단계에 중요한 영향을 미치는 핵심적 시스템으로서의 기능을 갖추어야 한다.

이를 위하여 사전탐색과 사후탐색에 있어서 적합성을 근거한 초기 질의어의 자동확장, 데이터베이스 선택과 탐색전략 구축에 있어서 불리안 로직과 매칭함수를 결합한 탐색식 구성과 탐색수행의 기능을 갖는 클러스터링을 이용한 시소러스 브라우저가 필요하다.

Ⅲ. 시소러스 브라우저 구축의 이론적 배경

1. 시소러스 자동구축 과정

시소러스를 구축하는 방법은 기존 시소러스의 활용 여부와 어휘 선정시 주제전문가의 간섭 정도에 따라 구분된다. 일반적으로 시소러스를 구축하는 방법은 다음과 같은 3가지 방법이 해당된다. 첫째, 동일 주제분야의 시소러스가 이미 만들어져 있는 경우, 이에 대한 최소한의 수정만을 통하여 구축한다. 둘째, 기존의 일반적 시소러스나 관련 분야의 시소러스, 또는 주제명표·분류표 등의 어휘집을 전체적인 틀로 사용하되 핵심주제의 용어는 별도로 수집하여 상세한 시소러스를 개발한다. 셋째, 용어를 새롭게 수집하여 완전히 새로운 체제의 시소러스를 개발한다.

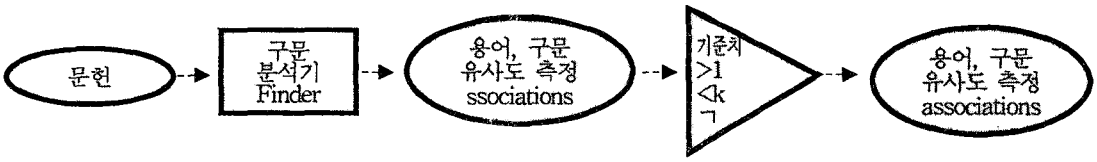
주제전문가의 간섭 정도에 따른 시소러스 구축방법은 주제전문가들의 합의를 통하여 구축하는 방법과 문헌에서 용어를 추출하여 구축하는 방법이 있다. 이에 대한 많은 선행연구에서 주제전문가들의 노동집약적인 합의에 의한 방법보다 문헌에서 용어를 추출하여 구축한 시소러스로 검색한 결과가 최신용어를 수용할 가능성이 높기 때문에 검색의 효율이 좋다는 결과를 발표하고 있다. 또한 시소러스를 구축하는데 많은 인력과 비용과 시간적 노력이 소요되므로, 이에 대한 해결책으로 문헌에 출현하고 있는 용어를 이용해 시소러스를 자동으로 구축하는 방법에 대한 관심이 높아지고 있다.

특히 시소러스에 출현하는 용어들은 문헌보증(Literary Warrant)과 이용자보증(User Warrant)

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

원칙을 따라야 한다.[15] 문헌보증은 서지적 보증(Bibliographic Warrant)이라고도 하며, 용어를 선정할 때 그 용어가 검색을 목적으로 중요하고 유용한 문헌에 충분히 자주 출현할 경우에만 정당화될 수 있다는 의미이며, 이용자보증은 용어의 특정성 수준을 적절하게 구축할 때에 특히 중요한 것으로 문헌에 나타나는 용어는 이용자가 사용하는 것보다 더 특정적일 수 있다는 의미이다.

따라서 본 연구에서는 시소러스의 문헌보증과 이용자보증 원칙에 따라 시소러스를 구축하는 방법으로 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 용어쌍(Term Pair)은 의미라는 측면에서 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 자동 구축방법

앞에서 설명한 시소러스 구축방법을 단계적으로 기술하면 다음과 같다.

첫 번째 단계는 어휘생성 단계로서 적절한 문헌을 수집하고, 시소러스의 특정성을 결정하고, 이를 근거로 대상 문헌의 서명, 초록, 원문을 대상으로 용어를 수집한다. 수집된 용어를 근거로 정규화과정을 거치는데, 이 과정은 전치사, 접속사와 같은 단어를 불용어(Stop Word)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, 이에 용어들의 동시출현 빈도를 고려한 cosine, dice 등의 유사성 측정공식을 적용해 문헌간의 유사성을 확인한다. 이 단계에서 포함된 모든 문헌들은 계층화를 이루도록 클러스터링 알고리즘을 적용하며, 이때 클러스터의 표현은 클러스터를 형성케 한 센트로이드에 해당하는 용어들로 표현한다.

세 번째 단계는 어휘의 관련도를 근거로 어휘들의 관계를 상위어, 하위어, 관련어로 표현하는

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

단계이다. 그 방법은 최상위레벨부터 시작해 인접레벨(Adjacent Level) 클러스터의 센트로이드를 비교해 처리하는 단계를 거친다. 먼저 직계계층인 부모-자식 형태로 이어지는 인접레벨 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 시멘틱 웹, 지식 데이터베이스 등의 구축에 기초적인 대안이 될 것이다.

2. 시소러스 자동구축 알고리즘

1) 자동색인 알고리즘

자동색인은 컴퓨터에 입력된 문헌을 대상으로 문헌의 내용을 나타낼 수 있는 단어나 단어구를 추출하는 과정이며 색인 과정에서 분석대상이 되는 부분은 문헌의 전문이나 초록이 된다. 컴퓨터에 의한 자동 색인은 시소러스 이용여부에 따라 시소러스 기반 색인법과 일반 색인 기법(단일어 색인 기법)으로 나누어진다.

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 연구의 색인작성법을 일반색인 기법(단일어 색인 기법)으로 한정한다.

일반 색인 기법은 색인어를 선정하는 기준에 따라 통계적 기법, 언어학적 기법, 문헌구조적 기법 등의 3가지 방법이 있다. 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 기법, 구문적 단계 기법, 어의적 단계 기법으로 나누어진다. 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 같이 주제를 잘 나태는 문장을 선택하여 이 문장 속에 나타난 주제어를 색인어로 선택하는 방법이 있다.

그러나 대부분의 한글 자동색인법은 언어학적 기법을 이용하여 색인의 대상이 되는 명사나 명사구를 식별하고, 통계적 기법을 이용하여 식별된 명사나 명사구를 색인어로 적용시키는 방법을 채택하고 있다.[16]

2) 클러스터링 알고리즘

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

클러스터를 구성하는 일반적인 과정은 다음과 같다. 먼저 문헌-색인어 행렬을 대상으로 유사치(Similarity) 측정 공식을 적용시켜 각 문헌쌍 { Di, Dj } (i = 1, 2,···, k / j = 1, 2,···, k) 들 간의 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 Structure)로 조직된다.[17][18]

계층적 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 않고 있다. 현재 주로 이용되고 있는

16) 남영준, 색인어형태분석에 의한 한국어 자동색인기법 연구(박사학위논문, 중앙대학교 대학원 문헌정보학과, 1994).

17) N. Jardine and C. J. Van Rijisbergen, "The Use of Hierachic Clustering in Information Retrieval," *Information Storage and Retrieval*, 7(1971), pp.217-226.

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

응집적 방법은 단일연결(Single Link), 완전연결(Complete Link), 그룹평균연결(Group Average), Ward방법(Ward's Method) 등이며, 기타 중앙값(Median)방법과 중심값(Centroid)방법이 있다.

3) 자동탐색 알고리즘

정보탐색과정의 핵심은 질의어 선정을 위한 사전탐색, 선정된 질의어들과 불리안 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 부족은 물론, 불리안 로직을 이용한 검색결과의 축소 및 확대에 필요한 탐색전략(Strategy)과 기법(Tactic)등에 대한 이해가 부족하다. 따라서 이용자를 대신해 검색 질의어 확장, 탐색전략 구축 및 탐색, 피드백 탐색 등의 과정을 수행하는 시스템이 필요하다.

질의어 확장은 효과적인 검색을 위해서 이용자의 초기 질의어에 포함된 탐색어에 이형동의어, 동의어, 관련어 등을 추가하는 과정으로써 자동탐색전략의 전단계에 해당된다. 질의어 확장방법에는 시소러스나 의미네트웍 등과 같은 정보원을 사용해 질의어를 확장하는 지식기반확장 방법[19]과 초기질의 벡터의 탐색으로 검색된 문헌 중 적합문헌에 출현한 용어들을 사용해 확장되는 탐색결과기반확장방법 등이 있다.[20]

탐색방법의 종류는 크게 불리안 탐색과 매칭 함수(Matching Functions)에 의한 탐색 방법으로 나뉘어진다. 불리안 탐색은 and, or, not 등의 연산자를 근거로 정보를 검색하는 방법이며, 매칭함수를 이용한 탐색은 질의어를 문헌이나 클러스터와의 연관성을 근거로, Dice 계수·cosine 계수·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.[21]

---

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

20) 노정순, "탐색결과에 근거한 자연어질의 자동확장 및 응용에 관한 연구 고찰," 정보관리학회지, 제16권, 제2호(1999), pp.49-80.

21) C. J. Van Rijisbergen, *The Hyper-Textbook of the C.J. Van Rijsgergen's Textbook on Information Retrieval*, 1998, <http://www.dei.unipd.it/~melo/bible/docu ments>.

피드백 탐색은 검색결과가 만족스럽지 못할 경우 새롭게 탐색을 수행하는 과정을 의미하는 것으로 탐색결과에 근거한 질의확장(query expansion search based on search results)이라고도 한다. 탐색결과를 근거로 자동으로 질의를 확장하여 정보를 탐색하는 방법은 매칭함수를 이용해 구성이 가능하다. 자동 피드백 탐색 방법은 먼저 초기 질의어를 이용해 검색된 문헌들에 대해 매칭함수를 적용한 적합성 서열화가 이루어져야 하며, 이를 근거로 가장 적합하다고 판단되는 10% 이내의 문헌에 출현하는 용어를 근거로 초기질의를 확장 또는 수정하여 탐색을 수정하는 과정을 거친다. 적합성 피드백에 의한 초기질의 수정 또는 확장 방법은 질의어 자동 수정 방법(Automatic Query Modification), 질의어 자동 확장 방법(Automatic Query Extension - All), 질의어 자동 선별 확장 방법(Automatic Query Extension - Sellect), 탐색자 개입 질의어 확장 방법(Interactive Query Expansion) 등이 있다.

Ⅳ. 시소러스 브라우저 자동구현을 위한 색인어 처리방안

1. Metadata를 이용한 색인어 처리

앞에서 기술한 바와 같이 시소러스를 구축하기 위해서는 특정 주제분야를 선택해서 그 분야에 해당하는 명사 형태의 용어를 생성하는 단계, 클러스터링 알고리즘을 이용해서 추출된 용어 간의 유사성 정도를 측정하는 단계, 측정된 유사성 정도를 근거로 상위어, 하위어, 관련어 등과 같이 용어들의 계층관계를 표현하는 단계 등을 거친다. 이와 같은 단계에서 첫 단계인 명사 형태의 용어를 생성하는 단계가 색인어를 추출하는 단계이다. 웹 문헌에서 색인어를 추출하는 방법은 수록되어 있는 메타데이터를 분석해서 주제를 의미하는 문장이나 주제어를 추출함에 의해 해결할 수 있다.

현재 웹 상에서 유통되고 있는 문헌 중 일부 문헌에는 메타데이터(Metadata)가 수록되어 있다. 대부분 DC(Dublin Core)메타데이터를 수록하고 있는데, 그 세부사항은 정보자원의 이름(서명, 영화명, 사진명, 음악명 등)에 해당하는 <meta name="DC.title">, 정보자원의 지적 내용에 대한 책임을 갖는 인물명, 단체명, email 주소(저자나 화가, 사진작가, 삽화가 등)를 나타내는 <meta name="DC.creator">, 정보자원의 주제나 그 내용을 표현한 명사나 명사구(자연어나 LCSH, DDC 등의 통제어휘나 분류기호)를 의미하는 <meta name="DC.subject">, 정보자원에 대한 문장(또는 명사구)형태의 설명을 기술(목차, 초록, 서론, 영화평 등)한 <meta name="DC.description">, 현재 형태의 자원을 이용 가능하게 만든 존재(출판사, 대학, 기업체 등)를 나타내는 <meta name="DC.publisher">, CREATOR에 포함된 인물 이외의 인물이나 단체(편자, 이기자, 삽화가 등)를 나타내는 <meta name="DC.contributor">, 현재 형태의 자원이 제작된 연도(YYYY-MM-DD, YYYY, 월 YYYY

등)를 나타내는 <meta name="DC.date">, 자원의 내용 형태(홈페이지, 시, 기술보고서, 논문, 사전, 소프트웨어, 사진 등)를 나타내는 <meta name="DC.type">, 자원의 표현방식(text/html, ASCII, Postscript file, 실행 가능한 응용프로그램, JPEG 등)을 기술한 <meta name="DC.format">, 자원을 고유하게 식별하기 위한 문자열이나 부호(URI, URL, DOI, ISBN)에 해당하는 <meta name="DC.identifier">, 해당 자원의 출처가 되는 인쇄나 전자형태의 저작(현 웹 자원에 대한 인쇄매체)형태를 나타내는 <meta name="DC.source">, 정보자원을 표현하는데 사용한 언어(한글, 영어 등 RFC 1766 정의 언어 값)를 나타내는 <meta name="DC.language">, 독립적으로 존재하는 자원간의 공식적인 관계(문서내의 그림, 도서의 일부 등)를 나타내는 <meta name="DC.relation">, 정보자원의 내용에서 다루어진 지역이나 시대(19세기 한국, 조선시대의 서울)를 나타내는 <meta name="DC.coverage">, 정보자원 전체에 대한 권리정보(copyright 소유권자, copyright 연결 URL이나 URI)를 나타내는 <meta name="DC.rights"> 등을 수록하고 있다.

이상과 같은 메타데이터에는 해당 문헌의 주제를 나타내는 테그가 있어 다양한 형태의 주제어 추출이 가능하다. 주제를 나타내는 테그에는 서명을 표현하는 곳인 <meta name="DC.title">, 주제어를 표현하는 곳인 <meta name="DC.subject">, 주제를 문장 형식으로 기술해 초록의 형태라 할 수 있는 <meta name="DC.description"> 등이 해당된다.

저자가 'Dave Beckett'이며 서명이 "Dave Beckett's Resource Description Framework (RDF) Resource Guide"인 웹주소 http://www.ilrt.bristol.ac.uk/discovery/rdf/resources/의 메타데이터 중에서 주제를 나타내는 테그를 분석하면 그 결과는 <그림 2>와 같다.

```
<meta name="DC.title" lang="en" content="Dave Beckett's Resource Description Framework (RDF) Resource Guide" />
<meta name="DC.subject" lang="en" content="Resource Description Framework; RDF; Dublin Core; metadata" />
<meta name="DC.description" lang="en" content="A page of resources on the Resource Description Framework (RDF) technology including examples, documents, software and links to projects." />
```

<그림 2> 주제 표현 메타데이터와 기술 내용

2. Metadata 자동편집기를 이용한 색인어의 처리

1) 수작업 처리방법

현재 연구 개발중인 메타데이터 편집기는 Reggie : The Metadata Editor, DC Dot, S-Link_S Editor/Publisher, RDF Schema editor, RDFPic, GraMToR, Protege, Mozilla, Metabrowser 등이

있다.[22]

본 연구에서는 DC Dot(http://www.ukoln.ac.uk/metadata/dcdot/)을 중심으로 기술하고자 한다. DC Dot을 이용하면 RDF, IEEE LOM, IMS, HTML, USMARC, SOIF, TEI Header, IAFA/ROADS, GILS, OLSTF, XML 형태의 메타데이터를 자동으로 구축할 수 있다.

DC Dot의 메타데이터 편집방법의 원리는 다음과 같다. ①주소입력창에 자신이 원하는 웹 사이트의 URI를 입력해 해당 웹페이지를 불러와 HTML 소스 내의 <head></head> 테그 안에 수록되어 있는 메타데이터를 분석한다. ②수록되어 있는 메타데이터를 DC 메타데이터의 15개 요소로 표현한다. ③DC 15개 요소에 따라 표현된 정보를 분석해 이용자가 필요한 정보를 추가하거나 변경 수정한다. ④수정변경된 메타데이터를 이용자의 필요(파일의 종류)에 따라 HTML, XHTML, XML, RDF의 메타데이터로 변경한다. ⑤원하는 형태로 변경된 메타데이터를 해당 정보자원의 메타데이터로 저장한다.

예를 들면 Andy Powell의 논문 "Expressing Dublin Core in HTML/XHTML meta and link elements"의 내용이 수록되어 있는 웹주소 'http://dublincore.org/documents/dcq-html/'를 입력하면 다양한 메타데이터들을 자동으로 구성할 수 있다. 앞의 웹주소에 대한 DC Dot에서의 HTML 메타데이터로의 자동 변환결과는 <그림 3>과 같다.

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/">
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/">
<meta name="DC.title" content="Expressing Dublin Core in HTML/XHTML meta and link elements">
<meta name="DC.creator" content="Andy Powell, UKOLN, University of Bath">
<meta name="DC.subject" content="Information; Dave; Reserved; title; recommended; follows; Simon;
appropriate; applications;
        ...생략...
; XML; Documents; management, public; wrapped; recommendation; mixed; sequence; Copyright; naming; Hors;
elements
<meta name="DC.type" scheme="DCTERMS.DCMIType" content="Text">
<meta name="DC.format" scheme="DCTERMS.IMT" content="text/html">
<meta name="DC.identifier" scheme="DCTERMS.URI"
 content="http://dublincore.org/documents/dcq-html/">
```

<그림 3> 수정된(Information이 추가된) HTML 메타데이터

<그림 3>의 데이터와 같이 DC의 Subject에 'Information'이란 단어를 새롭게 추가한다면, 이에 따라 새롭게 'Information'이란 단어가 추가된 새로운 <meta name="DC.subject">를 자동으로 구성할 수 있다. 이와 같이 DC Dot을 이용하면 해당 웹페이지에서 색인어 추출이 가능한 <meta name="DC.title">, <meta name="DC.subject">, <meta name="DC.description"> 등의 수정이 가

22) 신현성, 시맨틱 웹을 위한 RDF 편집기(석사학위논문, 경기대학교정보통신대학원, 2002), p.26.

능하다.

또한 특정 웹 문헌에 <meta name="DC.subject">가 없을 경우에는 해당 문헌을 직접 읽어보고 그 문헌을 대표하는 색인어들을 직접 입력함에 의해 <meta name="DC.subject">를 구성할 수 있다. 만약 해당 웹 문헌에 주제를 표현한 메타데이터인 <meta name="DC.title">, <meta name="DC.description"> 등이 있다면 이를 분석해서 <meta name="DC.subject">에 포함이 가능한 색인어를 수작업으로 추출해 입력할 수 있다.

그리고 DC Dot에서는 수정된 HTML 메타데이터를 이용해 XML 메타데이터와 RDF 메타데이터로의 자동 변경이 가능하다. 그 결과는 <그림 4>와 <그림 5>에 제시되어 있다. 자동으로 변경된 내용은 XML 파일은 .xml 형식으로, RDF 파일은 .rdf 형식으로 저장해 HTML 파일 내에 별도의 메타 테그를 사용하지 않고 <head> 테그 안에 <link rel="meta" href="index.html.rdf"> 구문을 추가해 입력하면, 메타데이터로서의 역할을 수행할 수 있다.

```
<?xml version="1.0"?>
<metadata
 xmlns="http://www.ukoln.ac.uk/metadata/dcdot/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.ukoln.ac.uk/metadata/dcdot/
  http://www.ukoln.ac.uk/metadata/dcdot/dcdot.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title>
   Expressing Dublin Core in HTML/XHTML
meta and link elements
  </dc:title>
  <dc:creator>
    Andy Powell, UKOLN, University of Bath
  </dc:creator>
  <dc:subject>
    Information; Dave; Reserved; title; recommended;
follows; Simon; appropriate; applications;
     ...생략...
  </dc:subject>
  <dc:description>
    automatic HTML metadata generator
  </dc:description>
  <dc:publisher>
    seo whee
  </dc:publisher>
  <dc:type>
    Text
  </dc:type>
  <dc:format>
    text/html
  </dc:format>
  <dc:identifier>
    http://dublincore.org/documents/dcq-html/
  </dc:identifier>
</metadata>
```

<그림 4> XML 메타데이터

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF SYSTEM
"http://dublincore.org/documents/2002/07/31/dcmes-xml/dc
mes-xml-dtd.dtd">
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description
  rdf:about="http://dublincore.org/documents/dcq-html/">
  <dc:title>
    좌동
    </dc:title>
    <dc:creator>
      Andy Powell, UKOLN, University of Bath
    </dc:creator>
    <dc:subject>
   Information; 좌동
    </dc:subject>
  <dc:description>
    automatic HTML metadata generator
  </dc:description>
  <dc:publisher>
    seo whee
  </dc:publisher>
  <dc:type>
    Text
  </dc:type>
  <dc:format>
    text/html
  </dc:format>
  </rdf:Description>
</rdf:RDF>
```

<그림 5> RDF 메타데이터

2) 자동색인방법

현재 가장 이용이 많이 되고 있는 웹 문헌은 전문(Full Text) 형태의 웹 문헌이다. 이들 웹 문헌에는 메타데이터가 수록된 것도 있고, 전혀 메타데이터를 수록하지 않은 것도 있다. 메타 데이터가 수록된 웹 문헌 중, <meta name="DC.subject">는 없고, <meta name="DC.title">과 <meta name="DC.description">만 있을 경우에는 이들 테그에 수록된 내용들을 자동색인 기법을 적용해 자동으로 <meta name="DC.subject">에 수록된 색인어들을 추출할 수 있다.

그러나 현재 대부분의 웹 문헌에는 최종 이용자의 무관심으로 인해 대부분 메타데이터를 수록하고 있지 않고 있다. 이를 해결하기 위해서는 웹 문헌의 첫 번째 절(Paragraph)을 대상으로 문장 형태인 <meta name="DC.description">을 해결하고, 이를 자동색인 기법을 적용함을 통해 <meta name="DC.subject"> 항목을 해결하는 방법이 도입되어야 할 것이다.

웹 문헌의 첫 번째 절을 대상으로 문장 형태인 <meta name="DC.description">을 해결한 후에 <meta name="DC.subject"> 항목에 수록될 색인어를 자동으로 추출하는 방법을 해결하기 위해서는 용어 분석이 필요하다. 용어 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.[23]

(1) 용어분석 알고리즘

문장 내에서 분리기호를 이용하여 어절을 분리하는 과정을 의미한다. 한글에서는 문자열과 문자열을 분리하는 식별자로써 공백문자(' ')를 사용한다. 그리고 생성된 문자열에서 <그림 6>에 제시된 기호들은 잘못된 문자로 인식하고 제외시킨다.

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

<그림 6> 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 비

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

교해 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 문제점이 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-gram 방법은 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.

V. 결론

본 연구에서는 인터넷에서 유통되고 있는 웹 문헌을 대상으로 자동으로 시소러스를 구현하는 방안에 대해 기술하였다. 특히 웹 문헌에서 그 문헌을 대표하는 주제어들을 수록하는 데이터가 있다면, 기존의 시소러스 구축방법을 이용해서 자동으로 시소러스를 구성할 수 있는 가능성에 대하여 연구하였다.

따라서 본 연구에서는 시소러스를 자동으로 구성하기 위한 이론적 연구의 핵심으로서 색인어 추출방법과 추출된 용어들간의 관계를 측정하여 용어들간의 계층을 자동으로 형성하는 클러스터링 알고리즘에 대해 분석하였다.

그리고 이들 알고리즘이 적용되기 위해서는 우선 색인어를 자동으로 추출하는 방법이 선행되어야 함에 착안하여 인터넷 상에 유통되고 있는 웹 문헌의 서지사항이라 할 수 있는 메타데이터를 분석하여 DC 메타데이터에서 서명, 초록, 주제어에 해당하는 항목들을 쉽게 해결할 수 있는 방안을 모색하였다.

그 결과 메타데이터 자동편집기를 이용해서 주제를 나타내는 메타데이터인 <meta name="DC.description">과 <meta name="DC.subject">를 쉽게 처리할 수 있는 방안을 제시하였다. 그 방법은 DC Dot 등과 같은 메타데이터 자동편집기를 이용해서 수작업으로 색인어를 입력해서 <meta name="DC.subject"> 항목을 자동으로 변환하는 방법임을 알 수 있었다. 또한 자동으로 <meta name="DC.subject"> 항목을 처리하기 위해서는 전문 내의 첫 번째 절을 초록이라고 판단하고 이를 <meta name="DC.description">으로 처리하는 방안을 제안하였다.

그러나 이와 같은 색인어 자동처리를 위한 방법보다 선결되어야 할 것은 시소러스와 온톨로지 그리고 시멘틱 웹을 구축하기 위해서는 먼저 인터넷의 특성 상 정보유통의 핵심적 역할을 하는 최종 이용자의 메타데이터에 대한 인식과 함께 이를 웹 문헌에 반드시 수용할 수 있는 방안이 모색되어야 할 것이다.

<참고문헌은 각주로 대신함>