질의어 확장에 기반을 둔 클러스터링 및 필터링 문서의 검색효율 제고에 관한 연구

A Study on the Improvement of Retrieval Effectiveness to Clustered and Filtered Document through Query Expansion

노 동 조(Dong-Jo Noh)*

초 록

인터넷을 비롯한 대다수의 정보검색에서 사용자가 느끼는 공통된 어려움중의 하나는 검색결과가 너무 많다는 것이다. 본 연구는 검색결과를 줄이는 방법의 하나로써 검색 문헌에 대한 정제 방법에 대하여 논의한 것이다. 궁극적으로 종전의 검색시스템에서 제대로 고려하지 않은 개념망을 통한 질의어 확장과 확장 질의어와 전처리된 문서와의 유사도 측정을 통한 문서의 선택, 백과사전 정보에 의한 의미 확장과 클러스터링, 필터링 기법 등이 정보검색의 효율을 향상시키는데 효과적인 방안임을 제안한다.

ABSTRACT

The purpose of this study is to improve of retrieval effectiveness to clustered and filtered document through query expansion. The result of this research prove that extended queries and documents, information in encyclopedia, clustering and filtering techniques are effective to promote retrieval effectiveness.

키워드: 분류, 클러스터링, 필터링, 질의어 확장, 정보검색, 검색효율
classification, document clustering, document filtering, query expansion, information retrieval, retrieval effectiveness

---

* 혜천대 IT정보계열 조교수(djnoh@hcc.ac.kr)
논문접수일자 2003년 6월 7일
게재확정일자 2003년 6월 20일

1. 서론

정보검색에서 사용자가 느끼는 공통된 어려움중의 하나는 검색된 문헌의 수에 있다. 어떤 경우는 검색된 문헌이 한 건도 없거나 너무 소수의 문헌만이 검색되어서 적합문헌의 수가 절대적으로 부족한 경우이고, 다른 하나는 검색된 문헌의 수가 과다하게 많은 경우이다.

그런데 최근 인터넷을 위시한 대부분의 정보검색에 검색자가 보다 빈번히 직면하는 문제는 후자의 경우이다. 이러한 상황에서 검색자는 우선적으로 2차 탐색을 통해서 검색문헌의 수를 줄이려고 할 것이다.

하지만 2차 탐색을 통해서도 검색문헌의 수를 줄이는데 효과를 보지 못하는 경우가 자주 발생하며, 더욱 난감한 상황은 검색문헌의 수를 줄이는 것에만 집착하여 너무 협소한 검색전략을 사용한 결과, 1차 탐색에서는 있었던 중요한 문헌이 사라지는 경우이다.

정보검색 분야에서는 앞서 지적한 문제들을 해결하기 위해서 꾸준히 연구가 진행되어 왔다. 현재 논의되고 있는 주요 방향은 두 가지로 압축된다. 하나는 사용자의 질의어를 기존의 키워드가 조합된 형태가 아닌 자연어 질의어처럼 개선된 형태로 제공하여 정보를 검색하는 방법이고, 다른 하나는 기존의 키워드 질의어 방식의 정보검색 방법을 통해서 얻은 결과를 수집, 분석, 가공하여 사용자의 요구에 맞는 정보만을 선별하여 제공하는 것이다.

본 연구는 위의 두 가지 방법 중에서 후자의 방법에 주목을 한다. 즉, 문서 필터링을 통하여 검색된 결과를 분류, 정제하고 사용자의 요구를 충족시켜주는 결과만을 제공하여 궁극적으로 정보검색의 효율을 향상시키기 위한 방법에 대해서 논의한다.

전술한 연구의 목적을 달성하기 위하여 검색결과의 정제에 관한 이론적 배경과 선행연구를 개관하고 나아가 질의어 확장에 기반을 둔 클러스터링 및 필터링 문서의 검색효율 향상방안에 대하여 제안한다.

2. 이론적 배경

2.1 문서 분류(Document Classification)

문서 분류는 일정한 분류 체계에 기초하여 분류의 대상이 되는 문서를 가장 적합한 범주(Category)나 클래스(Class)에 할당함으로써 문서들의 집단을 형성하는 방법이다. 문서분류를 이용한 대표적인 검색시스템으로는 WebWatcher와 NewsWeeder가 있는데 이 중 WebWatcher는 웹 환경에서 사용자에 적절한 탐색경로를 제공해 줌으로써 사용자가 원하는 정보를 쉽게 찾을 수 있도록 도와주는 시스템이며, NewsWeeder는 유즈넷 뉴스(Usenet News)에서 제공하는 방대한 양의 정보중에서 사용자의 관심사에 해당하는 문서만을 여과시켜주는 시스템이다.

문서 분류를 위해서는 문서를 특성 벡터(Feature Vector)의 형태로 표현하여 처리하며 특성 벡터의 구성은 단어의 빈도(Term Frequency)나 역문헌 빈도(Inversed Document frequency) 또는 위의 두 가지 방법을 적절히 결합한 형태로 이루어 진다. 그렇게 해서 얻어진 특성 벡터중에서 문서를 대표할 수 있는 주요 특

성만을 선택하여 사용한다.

문서빈도 제한기법은 특정 단어가 나타나는 문서의 수(Document Frequency)를 이용하는 방법이다. 이는 문서의 집합에서 각 단어마다 문서의 빈도를 구한 후, 미리 정의된 임계값(Threshold)보다 작은 빈도를 갖는 문서를 제거하는 방식이다.

2.2 문서 클러스터링(Document Clustering)

문서 클러스터링은 문서간의 유사성에 근거하여 유사한 문서들의 집단을 형성하는 방법이다. 클러스터링을 수행하기 위해서는 클러스터를 생성하고자 하는 문서의 집합에서 각 문서를 특성 벡터(Feature Vector)로 표현하고 주요 특성을 선택한다. 문서간의 유사도 측정을 위해서 기본 데이터인 문서-색인어 행렬을 사용하여 유사성 척도를 구하고 이를 클러스터링 알고리즘에 따라 문서를 분류한다.

클러스터링 알고리즘의 방식은 계층적기법과 비계층적기법으로 양분된다. 계층적 클러스터링 기법은 최상위에 모든 문서를 포함하는 하나의 클러스터에서부터 출발하여 최하위에 각 문서가 할당되는 모형으로 최하위의 클러스터로부터 더 큰 클러스터를 생성하기 위해서 클러스터 사이에 일정한 거리 값에 따라 단계적으로 결합을 반복한다. 계층적 클러스터링 기법으로는 단일 연결(Single Linkage), 완전 연결(Complete Linkage), 그룹 평균(Group Average), 워드 기법(Ward's Method), 중앙치 방법(Median) 등이 있다. 이 중에서 가장 일반적으로 사용하는 방식은 워드 기법으로 클러스터 중심간의 유클리드 거리(Euclidian Difference)를 최소화하는 방법으로 클러스터를 생성한다. 워드 기법에 의해 생성된 클러스터는 균등하고 대칭적인 특성을 갖는 경향이 있다.

비계층적 클러스터링기법은 클러스터의 계층을 고려하지 않고 각 문서를 평면적으로 클러스터링하는 방법이다. 이 방법은 미리 몇 개의 클러스터로 나뉘어 놓고, 문서와 클러스터 중심간의 유사도 측정을 통해서 문서를 클러스터에 재배치하는 방식이다. 비계층적 클러스터링기법으로는 K-means기법과 EM기법이 있다. K-means기법은 재배치 기법의 일종으로 임의의 문서 K개를 선택하여 초기 클러스터의 중심으로 할당한 다음, K개의 클러스터 중심이 바뀌지 않는 동안 나머지 문서와 클러스터 중심간의 유사도가 가장 높은 문서를 클러스터에 포함시키는 방법이다. EM기법은 K-means기법의 변형으로, 첫 번째 클러스터링 작업에서 K개의 초기 중심 문서가 나머지 문서의 중심으로 이동하고 두 번째 반복에서 초기 클러스터를 생성하는 방법이다.

2.3 문서 필터링(Document Filtering)

문서 필터링은 1958년 룬(Luhn)의 Business Intelligence System에서 처음 소개되었다. 룬의 생각은 도서관에서 사용자 개개인에 대한 프로파일을 만들고 이 프로파일에 따라서 사용자 개인에 맞는 문서 목록을 만들자는 것이었다. 그의 아이디어는 정보의 선택 모듈이 갖는 기능을 선택적으로 설명함으로써 훗날 SDI(Selective Dissemination of Information)서비스에 관한 관심을 확산시키는데 큰 기여를 하였다.

필터링이란 용어는 대닝(Denning)이 CACM

의 1982년 3월호에 게재한 ACM Presidents Letter에서 공식적으로 처음 등장하였다. 여기서 그는 전자우편중에서 중요하고 시급한 것들을 구분하기 위해서 필터링이 요구된다고 역설하며, 구체적으로 컨텐츠 필터링 방법을 제안하였다.

말론(Malone)은 1987년 동료들과의 연구에서 필터링에는 세가지의 패러다임이 존재한다고 하였다. 첫째, 인지적 필터링(Cognitive Filtering)은 대닝이 처음 제안하였던 컨텐츠 필터링과 동일한 것으로서 오늘날의 내용기반 필터링(Content-based Filtering)을 의미한다. 둘째, 필터링에 관한 경제적 접근은 사용자 프로파일을 만드는데 따르는 비용을 고려한 것이며 셋째, 사회적 접근은 문서의 표현이 이용자들에 의해 만들어진 주석(Annotation) 정보에 기초하는 방식으로 현재의 협업 필터링(Collaborative Filtering)을 의미한다.

필터링을 위해서는 정보검색에 관한 기술과 필터링의 대상이 되는 동적 정보(Dynamic Information)에 대한 분석 및 관리 기술, 사용자의 관심을 표현하는 사용자 프로파일의 기술 등이 요구된다.

3. 선행연구

문서 클러스터링에 관한 초기의 연구는 클러스터 파일을 대상으로 검색 실험을 하기 위한 것이었으나 이후에는 유사한 환경에서 검색성능을 향상시킬 목적으로 여러 클러스터링 알고리즘을 비교 검토하는 연구로 발전하였다.

최근에는 검색대상 문헌을 클러스터 파일로 조직하기 위한 연구보다는 검색한 결과를 브라우징하거나 더 나아가 검색 결과를 분류하는 방법에 더욱 초점을 맞추고 있다.

대표적인 연구로는 Scatter/Gather(Cutting et al. 1992 ; Cutting, Karger, and Pederson 1993)를 들 수 있다. 이 연구는 데이터베이스의 내용을 사용자가 능동적으로 브라우징함으로써 사용자 스스로 적합 문헌을 찾아낼 수 있도록 하는 방식이다. 즉, 유사한 주제의 문헌들을 소집단으로 클러스터링한 다음, 각 클러스터의 요약 정보를 이용자에게 제공함으로써 몇 개의 선택된 클러스터만을 대상으로 다시 클러스터링 작업을 반복하여 종국에는 정보요구에 가장 적합한 클러스터를 생성하는 방식이다.

결국 문헌 클러스터링에 관한 연구는 검색 이전에 데이터베이스의 내용을 시각적으로 보여주기 위한 연구에서 출발했으나 이후에는 검색결과를 보여주거나 브라우징할 수 있도록 하기 위한 연구로 발전하였으며, 최근에는 검색된 문헌들의 클러스터링을 통해서 검색결과를 자동적으로 정렬하여 보여 줌으로써 검색 성능을 향상시키기 위한 연구로 발전하였다. 본 연구와 관련된 선행연구들을 개관한다.

정영미와 이재윤은 문헌을 기반으로 한 지식분류의 자동화를 위한 최적의 클러스터링 모형을 개발하고자 하였다. 이를 위해 신문기사 집단과 학술논문 초록 집단을 실험대상으로 하여 분류성능 평가 척도인 WACS(Weighted Average Cluster Similarity)를 개발하였다. 이 연구에서 분류자질로 사용한 용어의 집합은 다양한 자질축소 기준을 적용하였으며, 유사도 측정은 코사인 계수(Cosine Coefficients)와 자카드 계수(Jaccard Coefficients)를 적용하

였으며, 클러스터링 알고리즘으로는 계층적인 기법인 완전연결기법(Complete Linkage)과 비계층적 기법인 K-means 기법을 각각 병용하였다. 연구 결과, 신문기사 집단이 학술논문 초록 집단에 비해서 성능이 좋았으며, 계층적 기법인 완전연결기법이 비계층적 기법인 K-means기법보다 성능이 좋은 것으로 나타났다. 역문헌빈도의 적용은 완전연결 클러스터링에서는 긍정적인 효과를 나타냈으나, K-means 클러스터링에서는 그렇지 못한 것으로 나타났다. 분류자질은 전체의 7.66%만 사용하였을 경우에도 성능의 저하가 크지 않았으며, K-means 클러스터링에서는 오히려 성능 향상의 효과가 있는 것으로 나타났다.

정영미와 최상희는 문장 클러스터로부터 대표 문장을 선정하여 요약문을 자동으로 생성하는 자동요약 모형을 제시하고, 학습문서 집단을 이용하여 최적의 요약 환경을 구축하고자 하는 실험을 하였다. 문장의 클러스터링 기법으로 센트로이드 기법이 선택되었고 각 클러스터를 대표하는 문장의 선정을 위해서는 용어 및 문장의 가중치를 합산한 문장 값 기준이, 클러스터와 문장 벡터간 유사도 사이의 기준을 비교한 결과에서는 문장 값 기준이 선택되었다. 용어의 가중치로는 역문장빈도와 표제어 가중치, 그리고 문장의 위치 가중치가 자동요약 성능을 개선시키는 것으로 나타났다. 또, 적절한 요약문의 길이는 전제 문서의 1/3인 것으로 나타났다. 문서의 길이와 특성에 있어서 서로 성격이 다른 신문과 잡지기사의 두 집단을 대상으로 요약 실험한 결과, 요약 정확률은 신문기사 집단에서는 53%, 잡지기사 집단에서는 47%인 것으로 나타났다. 두 실험 모두 무작위로 생성한 베이스라인 요약문보다 성능이 우수하였으나, 리드(Lead) 문장들로 구성된 베이스라인 요약문과의 비교에서는 짧은 길이의 신문기사의 경우는 요약 모형의 성능이 오히려 떨어지는 것으로 나타났다.

김혜진과 문성빈은 하이퍼 텍스트나 웹 문서의 검색에서 링크로 연결된 문서들이 주제적으로 서로 연관되어 있다는 사실에 착안하여 링크 정보를 참조한 웹 문서 클러스터링 기법을 제안하고, 검색된 결과를 질의 근접 순위화함으로써 웹 문서 검색의 성능을 향상시키는 연구를 하였다. 연구에 사용한 웹 문서의 집단은 웹을 통하여 직접 수집하였으며, 웹 문서가 다른 웹 문서를 링크하고 있을 때는 OutLink, 다른 웹 문서로부터 링크받고 있을 때는 InLink로 구분하여 실험한 결과, OutLink를 참조하여 클러스터링을 수행하는 방법과 InLink를 참조하여 클러스터링을 수행하는 방법 모두에서 검색 성능의 향상을 보였다.

심지영과 김태수는 내용기반 음악검색 시스템에서 이용자의 탐색을 용이하게 하기 위해서 유사한 음렬을 검색 결과로 제시하였으며, 음렬 패턴을 대상으로 분류 자질을 선정하고 이를 기준으로 음렬간 유사도를 측정한 다음, 음렬간 군집을 형성하였다. 『A Dictionary of Musical Theme』에 수록된 주제 소절의 kern형식의 파일을 실험집단으로 하여 음렬의 분절 여부와 시작 위치에 따른 네가지 형태의 유사도 행렬을 대상으로 계층적 클러스터링 기법을 사용하였고 외적 기준이 되는 수작업 분류표가 있는 경우에는 WACS 척도를, 음렬내 임의의 위치에서부터 시작한 음렬을 대상으로 한 경우에는 클러스터링 결과로 얻은 군집내 공통 자질 패턴 분포를 통해서 내적 기준을 마련하여

평가하였다. 평가 결과, 음렬의 시작 위치와 무관하게 분절한 자질을 사용하여 클러스터링한 결과가 그렇지 않은 것에 비해 높게 나타난 것으로 밝혀졌다.

서휘는 전문정보를 대상으로 한 정보검색의 문제점이 정보탐색자의 탐색전략이나 탐색기법에 대한 인식 부족, 질의어의 표현·생성·확장의 어려움에서 발생한다고 보고 일련의 연구를 수행하였다. 이용자의 정보탐색 행태, 자동색인 작성법, 클러스터링 기법, 시소러스 구축 및 구현방법, 정보검색기법에 대해서 분석한 결과, 전문 데이터베이스를 대상으로 한 정보검색에서 검색효율을 극대화하기 위한 방법으로 클러스터링을 이용한 시소러스 브라우저 모형을 제시하였다.

정영미와 이용구는 최근에 주목을 받고 있는 협업 필터링기법을 중심으로 한 대출대상 도서의 추천 시스템을 실제로 구축하였다. 구체적으로 연관성 규칙 기반 기법, 협업 필터링 기법, 내용기반 필터링 기법을 응용하여 실제 대학도서관에서 측정 이용자가 대출할 만한 도서를 추천하는 시스템을 구현하고 각 기법의 추천 성능을 평가하였다. 연구 결과, 대출대상 도서를 추천하는데 있어 협업 필터링 기법과 내용기반 필터링 기법을 각각 따로 적용하는 것보다 두 기법을 함께 이용한 혼합한 필터링 추천 기법이 더욱 효과적인 것으로 나타났다.

이 외에도 백준호와 최준혁, 이정현은 정보검색의 효율 향상을 위한 웹 문서 필터링 에이전트 시스템 설계에 관한 연구를 수행하였으며, 박영우와 이은석은 인터넷 정보검색을 위한 사용자 적응형 필터링 방법에 관한 연구를 수행하였다.

선행연구를 통해서 살펴본 바와 같이 지금까지 문서 정제 즉, 클러스터링과 필터링에 관한 연구들은 대부분 여러 가지 클러스터링, 필터링 기법들을 사용하여 실험을 하고 그 결과가 검색효율에 미치는 영향을 규명한 연구가 대부분이었다. 그러나 보다 근원적인 문제는 문서를 추출하고 정제하는 과정에서 검색효율에 직접적인 영향을 미치는 질의어의 확장 및 절제에 관해서는 별로 연구가 이루어지지 않았다는 사실이다.

이에 본 연구자는 지금까지 이 분야의 연구가 기계적인 방법론에 치우쳐 검색효율에 직접적인 영향을 미치는 용어에 대한 통제나 확장에 관한 부분을 간과한 점에 문제를 제기하고 개념망을 통한 질의어 확장, 확장된 질의어 문서간의 유사도 측정을 통한 문서의 선택, 백과사전의 정의와 의미를 통한 필터링과 클러스터링 등의 방법을 통하여 검색효율을 향상시키는 방법에 관한 연구를 수행한다.

4. 질의어 확장을 통한 문서 필터링

4.1 개념망을 통한 질의어 확장

개념망이란 명사 어휘로 표현되는 개념을 보다 정확하게 정의하기 위해서 개념들간의 다양한 관계를 연결시켜 놓은 어휘 데이터베이스를 말한다. 개념망은 언어학적인 의미관계를 이용하여 상하관계를 기본 축으로 하고 있으며, 상하관계의 보완으로 개념들간의 동의 및 유의관계, 부분과 전체의 관계, 반의관계 등을 추가로 정의하고 있다.

본 연구에서 제안하고자 하는 개념망은 종전의 시소러스가 갖는 한계를 극복하고 다양한 측

면에서 개념들간의 관계를 포괄적으로 정리한 것이다. 즉, 종전의 시소러스는 개념들간의 관계의 모호성으로 인하여 용어들간의 상하관계가 명확하게 설정되지 않았으며 이러한 이유로 인해 상하관계를 통해서 공통된 질의 패턴과 속성을 추출하기 어렵다는 근본적인 문제점을 지니고 있었다.

예를 들면, 『문헌정보학』은 『인문과학』, 『사회과학』, 『학문』, 『지식』 등과 같은 단계적이면서도 다양한 의미를 지니고 있기 때문에, 기존의 시소러스에서도 이러한 점들을 감안하여 상하관계를 설정하고 있다. 그러나 종전의 시소러스 계층 구조는 대부분 개념어가 바로 위 단계의 상위어까지만이 이용이 가능하며 또, 공통된 속성을 형성하기 어려울 뿐만 아니라 자연어 처리기술을 이용한 정보검색에 있어서도 높은 효율을 기대하기 어려웠다. 하지만 본 연구에서 제시하는 개념망을 통해서 『문헌정보학』의 의미를 계층 구조의 차원에서 이해하고 구축하여 『문헌정보학-인문과학-학문-지식』 또는 『문헌정보학-사회과학-학문-지식』과 같은 계층적이면서도 논리적인 상하관계를 명확히 설정하여 줌으로써, 대치가능성을 고려한 질의 분석과 공통 속성의 연계성을 획득할 수 있다는 장점이 있다.

본 연구에서 제시하는 개념망의 또 다른 특징은 개념망 사전을 통한 질의어의 확장이 가능하여 검색된 문헌의 정제는 물론 누락된 정보에 대한 복원이 가능하다는 점이다. 특히 복합명사나 준말의 경우, 개념망 사전을 활용한 질의어의 확장은 검색효율을 높이는데 크게 기여할 것으로 보인다. 이런 경우에 확장 질의어는 질의어인 복합명사의 분해 결과와 질의어에 대한 상위어, 하위어 들로 구성된다. 만약, 질의어가 분해되지 않는 경우에는 상위어를 이용하여 질의어를 분할한다.

예를 들면, 『문서관리』라는 질의어가 있을 경우, 『문서관리』의 상위어는 『관리』이다. 이때, 질의어에서 상위어를 매칭하고 제거한 나머지 단어(문서)와 상위어(관리)가 확장 질의어에 포함된다. 또 다른 예로, 질의어가 『산업재산권』일 경우에는 상위어는 『권리』이다. 앞의 『문서관리』의 경우는 상위어가 질의어의 일부분으로 포함되지만 『산업재산권』의 경우는 그렇지 않다. 이러한 경우에는 『산업재산권』이 『산업재산권리』의 축약형이므로 이것을 개념망 사전을 통해서 유추, 해석하여 질의어의 마지막 음절과 상위어의 첫 음절을 비교하여 제거한 나머지 단어들(산업, 재산, 산업재산)과 상위어(권리)가 확장 질의어가 된다. 이처럼 질의어를 상위어에 따라 분리하는 근거는 상위어에 의해서 분리된 나머지 단어들(문서, 산업, 재산, 산업재산)이 문서의 일반적인 범주를 특징지어 주고, 그러한 정보는 문서 분류에서 매우 유용하기 때문이다.

4.2 유사도 측정을 통한 문서의 선택

통상적으로 질의어로 사용되는 용어들간의 유사 관계는 유사성(Similarity), 연관성(Association), 비 유사성(Dissimilarity)의 세가지 측면에서 측정된다. 이 중에서, 비 유사성은 단순히 용어의 유사성을 측정하는 수학적 방법에 해당되므로 용어들간의 관계를 표현하는 데에는 적합하지 않다. 일반적으로 연관성이란 용어를 더 많이 사용하지만 유사성과 연관성은 통계적으로는 동일한 의미를 갖는다.

용어들간의 유사도 측정은 정보검색의 질의

어를 확장하고자 할 때, 그 용어를 선정하기 위해서 반드시 필요하다. 유사도는 질의어로 사용하는 하나의 용어가 다른 용어와 어느 정도의 밀접한 관계에 있는 가를 측정하는 것이다. 유사도 측정에는 전통적인 시소러스를 이용하는 방법과 Corpus를 이용한 통계 정보를 이용하는 두가지 방식이 있다. 전통적인 시소러스를 이용한 용어간 유사도 측정은 두개의 용어 사이에 몇 개의 용어가 연결되어 있는 가를 계산하는 방식이며, 측정과 선택은 전적으로 시소러스 구축자의 주관에 의존한다.

Corpus 통계 정보를 이용하는 방법은 시소러스 방식과는 달리 두개의 용어가 동시에 출현한 공기 정보(Collation, Co-occurrence Information) 또는 상호 정보(Mutual Information) 등의 통계 정보를 통해서 유사도를 측정하는 방식이다.

용어들간의 유사도를 계산하는 공식에는 Dice 공식, Jaccard 공식, Cosine 공식, Overlap 공식, Tanimoto 공식 등이 있다. 이 중에서, Dice 공식과 Jaccard 공식은 분모에 해당하는 계수를 동시에 출현한 용어를 각기 1로 계산하는 방식(결국 합쳐서 2가 됨)이다. Overlap 공식은 분모를 적은 어휘를 동시에 출현한 문서의 어휘 수로 계산한 것이며, Tanimoto 공식은 분모를 동시에 출현한 용어를 1로 계산하는 방식이다.

본 연구에서 제안하고자 하는 유사도 측정 방식과 공식은 전술한 여러 방법중에서 가장 과학적인 방법인 Corpus를 이용한 통계 정보를 이용하는 방식을 채택하며, 확장 질의어와 문서간의 유사도를 계산하는 공식은 그 특성을 감안하여 아래의 공식을 적용할 것을 제안한다.

$$S_{di} = \frac{\Sigma(q_w \times q_{tf})}{\sqrt{TF_{di}}}$$

S_{di} : 문서 di와 확장 질의어와의 유사도
q_w : 확장 질의어 가중치
q_{tf} : 확장 질의어 빈도수
TF_{di} : 문서 di의 전체 명사의 수

위에서 보는 바와 같이 확장 질의어와 문서간의 유사도는 각 문서에 나타난 명사 가중치의 합을 그 문서의 전체 명사 수의 제곱근으로 나누어 유사도를 계산한다. 그렇게 해서 얻어진 유사도에 따라서 문서를 순위화하며, 사용자는 개개 문서의 유사도에 따라서 문서의 선택여부 결정한다.

4.3 백과사전을 통한 질의어 확장

문서 필터링의 결과, 질의어에 적합한 문서만을 선택해 낼 수 있다면 가장 이상적이다. 하지만 대부분의 문서 필터링의 결과는 그러하지 못하다. 본 연구에서 제안한 확장 질의어와 유사도 측정을 통한 문서의 선택 과정을 거치면 검색된 문서를 보다 명확하게 분류, 정제할 수 있다. 즉, 어느 정도의 재현율은 향상시킬 수 있으나, 그렇게 확장을 하다 보면 질의어에 적합하지 않은 문서도 많이 포함되기 때문에 결국 사용자에게 정확한 결과를 제공하는 데에는 한계가 따르기 마련이다.

따라서 본 연구에서는 바로 이러한 문제 즉, 재현율 향상을 위해서 질의어를 확장하고 유사도 측정을 통해서 문서를 선택하는 방식에 의해 발생하는 낮은 정확률의 문제를 백과사전에 있는 정보를 차용해서 해결할 것을 제안한다. 즉,

낮은 정확률의 문제를 백과사전의 정의와 의미를 되새겨 질의어를 확장하고 필터링된 문서중에서 적합하지 않은 문서를 제거하는 것이다.

이 방법은 백과사전에서 질의어에 대한 정의와 질의어에 대해서 자세하게 설명이 되어 있는 부분에서 질의어를 차용한 후, 여기서 필요한 명사들을 추출하는 것이다. 이 방법을 통해서 추출한 명사가 필터링된 문서에서 임계값 미만의 포함 비율을 갖는 문서들을 제거하면 자연적으로 정확률을 향상시킬 수 있다.

예를 들면, 『산업재산권』을 질의어로 했을 경우에 표 1은 질의어에 대한 백과사전의 정의와 내용을 나타낸 것이다. 표 2는 백과사전 정보에서 얻은 명사 즉, 확장 질의어를 표시한 것이다. 표 1과 표 2에서 보는 바와 같이 질의어 『산업재산권』에 대한 백과사전의 정의와 내용을 통해서 『산업』, 『재산』, 『재산권』, 『저작권』, 『무체재산권』, 『특허』, 『특허권』, 『실용신안권』, 『의장권』, 『상표권』, 『서비스권』 등이 확장 질의어로 채택되었고, 이들 질의어가 필터링된 문서에서 임계값 미만의 비율을 갖는 문서들을 제거하고 남은 문서들만을 선택하면 되는 것이다.

5. 질의어 확장을 통한 필터링 문서의 검색효율 향상

본 연구에서 제안한 백과사전 정보의 차용을 통한 정확률의 향상 방법은 필터링된 문서중에서 백과사전 정보의 포함 비율에 따라 문서를 제거하므로 정확률은 높일 수 있지만, 반대로 재현율이 다시 낮아진다는 문제점을 드러낸다. 이러한 단점을 보완하기 위해서 본 연구에서 문서 정제의 마지막 단계로 클러스터링을 활용할

표 1. 질의어에 대한 백과사전의 정의와 내용

질의어	백과사전의 정의	백과사전의 내용
산업재산권	산업상 이용가치를 갖는 발명 등에 관한 권리	산업재산권은 산업영역에의 기여에 대한 보호를 본질로 하며, 문화영역에 대한 보호를 본질을 하는 저작권(著作權)과 더불어 무체재산권(無體財産權)을 이룬다. 산업재산권은 좁은 의미에서는 특허권(特許權), 실용신안권(實用新案權), 의장권(意匠權), 상표권(商標權) 및 서비스표권(標權)을 말하며, 넓은 의미에서는 노하우(Know-How)권, 미등록주지상표권(未登錄周知商標權) 등 산업상 보호 가치가 있는 권리를 모두 포함하여 말한다. 산업재산권은 특허청(特許廳)에 등록을..... 〈이하생략〉

(출처 : www.naver.com)

표 2. 백과사전 정보에서 추출한 명사

질의어	백과사전에서 추출한 명사
산업재산권	산업, 재산, 재산권, 저작권, 무체재산권, 특허, 특허권, 실용신안권 .....

것을 제안한다.

본 연구에서 제시하는 클러스터링 기법은 백과사전 정보를 이용하여 제거된 문서의 수만큼 클러스터링을 통하여 문서를 다시 선택하는 것이다. 구체적인 방법은 앞서 본 연구의 4.2 유사도 측정을 통한 문서의 선택에서 이미 제시한 방법과 공식에 따라서 문서의 유사도를 측정하고 순위화된 문서중에서 높은 유사도를 갖는 상위 n% 문서의 중심 벡터를 생성한 다음, 중심 벡터와의 유클리드 거리(Euclidian Difference)가 가장 가까운 문서를 선택하여 필터링된 문서 그룹에 포함시키는 방식이다. 이렇게 함으로써 낮아진 재현율을 복원하고 정확률을 향상시킬 수 있다.

지금까지 검색문헌의 효율 향상을 위해서 본 연구에서 제안한 방식을 정리하면 다음과 같다. 먼저 개념망을 통한 질의어 확장을 통하여 재현율을 향상시킨다. 다음은 확장 질의어와 문서간의 유사도 측정을 통해서 일정한 문서를 걸러내어 정확률도 같이 높인다. 여기에서도 만족할 만한 결과를 얻지 못하면, 다음 단계로 백과사전의 정의와 내용을 통해서 다시 질의어를 확장한다. 마지막 단계는 백과사전 정보를 이용하여 클러스터링을 통해서 제거된 문서의 수만큼 문서를 재선택하는 것이다

6. 결론 및 제언

인터넷을 위시한 대다수의 정보검색에서 사용자가 느끼는 공통된 어려움 중의 하나는 검색결과가 너무 많다는 것이다. 본 연구는 검색결과를 줄이는 방법의 하나로써 사용자가 요구하는 정보만을 정제하여 궁극적으로 정보검색의 효율을 향상시키기 위한 방법에 대해서 구체적으로 제안하였다. 즉, 종전의 검색시스템에서 고려치 않은 개념망을 이용한 질의어 확장과 확장 질의어와 전처리된 문서와의 유사도를 측정하여 그에 따라 서 문서를 선택하고 백과사전 정보에 의한 의미 확장과 클러스터링 기법을 통하여 정보검색의 효율을 제고시킬 수 있는 방안에 대해서 논의하였다.

본 연구를 통해서 얻어진 결론을 요약하면 다음과 같다.

첫째, 개념망 사전은 개념들간의 다양한 관계를 연결시켜 놓은 어휘 데이터베이스로써 언어학적인 의미관계를 이용하여 상하관계를 기본 축으로 하여 개념들간의 동의 및 유의관계, 부분과 전체의 관계, 반의관계 등을 추가로 정의하고 있어서 종전의 시소러스가 갖는 관계의 모호성으로 인하여 발생하는 문제들을 극복할 수 있다.

둘째, 확장 질의어와 문서의 유사도 측정에 의해서 어느 정도의 재현율은 향상시킬 수 있으나 반대로 정확률은 떨어진다.

세째, 질의어 확장과 유사도 측정에 의한 문서의 선택을 통해서 발생하는 낮은 정확률의 문제는 백과사전의 정보를 차용하여 필터링된 문서중에서 질의어에 적합하지 않은 문서를 제거하면 된다.

네째, 백과사전 정보를 이용하여 정확률을 향상시키는 방법은 필터링된 문서중에서 백과사전 정보의 포함 비율에 따라 문서를 제거하므로 정확률은 높일 수 있지만, 반대로 재현율은 낮아진다. 이때 발생하는 낮은 재현율의 문제는 백과사전 정보를 이용하여 제거된 문서의 수 만큼 클러스터링을 통하여 문서를 다시 선택하면 된다.

본 연구자의 후속 과제는 이 연구에서 제안한 방식을 실제의 시스템에 적용하여 그 효용성을 검증하는 것이며 이는 곧 이어질 것이다. 나아가 문서 정제에 관심을 갖고 있는 후속 연구자들을 위해서 제안하면, 보다 성능이 높은 클러스터링과 필터링을 위해서는 질의어에 대한 가중치 할당의 문제와 함께 오류 문서의 유형을 분석하여 이를 개선하는 연구도 병행되어야 함을 밝혀 둔다.

참 고 문 헌

김혜진, 문성빈. 2002. 링크기반 클러스터링을 이용한 웹 문서 검색의 성능 향상에 관한 실험적 연구. 『제9회 한국정보관리학회 학술대회논문집』, 247-252.

노정순. 1999. 탐색결과에 근거한 자연어질의 자동확장 및 응용에 관한 연구 고찰. 『정보관리학회지』, 16(2): 49-80.

서 휘. 1999. 클러스터링을 이용한 시소러스 브라우저의 설계에 대한 이론적 연구. 『한국도서관·정보학회지』, 30(3): 427-456.

박영우, 이은석. 1998. 인터넷상에서의 정보검색을 위한 사용자 적응형 필터링 방법에 관한 연구. 『성균관대학교 논문집-과학기술』, 49(2): 63-74.

백준호, 최준혁, 이정현. 1999. 정보검색 효율 향상을 위한 웹 문서 필터링 에이전트 시스템 설계. 『산업과학기술연구소 논문집』, 27:511-516.

심지영, 김태수. 2002. 음렬 탐색을 위한 주제 소절 자동분류에 관한 연구. 『정보관리학회지』, 19(3): 5-30.

장문수 외. 2000. 인터넷 질의응답을 위한 지식베이스 구축. 『제12회 한글 및 한국어 정보처리학회 학술대회발표논집』, 198-204.

정영미, 이용구. 2002. 필터링 기법을 이용한 도서 추천 시스템 구축. 『정보관리연구』, 33(1): 1-17.

정영미, 이재윤. 2001. 지식 분류의 자동화를 위한 클러스터링 모형 연구. 『정보관리학회지』, 18(2): 203-230.

정영미, 최상희. 2001. 문장 클러스터링에 기반한 자동요약 모형. 『정보관리학회지』, 18(3): 159-177.

Armstrong, R. and Jocahims, T. 1995. WebWatcher: A Learning Apprentice for the World Wide Web. AAAI Spring Symposium on Information Gathering from Heterogeneous. Distribute Environment.

Cutting, D. et al. 1992. Scatter/Gather : A Cluster-based Approach to Browsing Large Document Collections. In Proceedings of the Fifteenth Annual ACM SIGIR Conference on Research and

Development in Information Retrieval. 318-329.

Cutting, D. et al. 1993. Constant Interaction-time Scatter/Gather Browsing of very Large Document Collections. In Proceedings of the Sixteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval. 126-134.

Joachims, T. and Mitchell, T. 1997. WebWatcher : A Tour Guide for the World Wide Web. International Joint Conference on Artificial Intelligence.

Lang, K. 1995. NewsWeeder : Learning to Filter News. International Conference on Machine Learning.

Larsen, B. and Aone, C. 1999. Fast and Effective Text Mining Using linear-time Document Clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Date Mining. 16-22.

Lee Kyung-Soon et al. 2001. Re-ranking Model Based on Document Clusters, Information Processing and Management. 37.

Schutze, H. and Silverstein, C. 1997. A Comparison of Projections for Efficient Document Clustering. In Proceedings of the Twentieth Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 74-81.

Vaithyanathan, S. and Dom, B. 1999. Generalized Model Selec-tion for Unsupervised Learning in High Dimensions. In Proceedings of the Neural Information Processing System.

Wong, Wai Chiu and Fu, A. 2000. Incremental Document Clustering for Wed Page Classification. In Proceedings of the IEEE 2000 International Conference on Information Society in the 21st Century : Emerging Technologies and New Challenges IS2000.