텍스트 마이닝 기법을 이용한 연관용어 선정에 관한 실험적 연구

An Experimental Study on Selecting Association Terms Using Text Mining Techniques

김수연(Su-Yeon Kim)*, 정영미(Young-Mee Chung)**

초 록

이 연구에서는 전체 문헌집단으로부터 초기 질의어에 대한 연관용어 선정 시 사용할 수 있는 최적의 기법을 찾기 위해 연관규칙 마이닝과 용어 클러스터링 기법을 이용하여 연관용어 선정 실험을 수행하였다. 연관규칙 마이닝 기법에서는 Apriori 알고리즘을 사용하였으며, 용어 클러스터링 기법에서는 연관성 척도로 GSS 계수, 자카드계수, 코사인계수, 소칼 & 스니스 5, 상호정보량을 사용하였다. 성능평가 척도로는 연관용어 정확률과 연관용어 일치율을 사용하였으며, 실험결과 Apriori 알고리즘과 GSS 계수가 가장 좋은 성능을 나타냈다.

ABSTRACT

In this study, experiments for selection of association terms were conducted in order to discover the optimum method in selecting additional terms that are related to an initial query term. Association term sets were generated by using support, confidence, and lift measures of the Apriori algorithm, and also by using the similarity measures such as GSS, Jaccard coefficient, cosine coefficient, and Sokal & Sneath 5, and mutual information. In performance evaluation of term selection methods, precision of association terms as well as the overlap ratio of association terms and relevant documents' indexing terms were used. It was found that Apriori algorithm and GSS achieved the highest level of performances.

키워드 : 텍스트 마이닝, 연관용어, 유사계수, Apriori 알고리즘, 용어 클러스터링
text mining, association terms, similarity measures, Apriori algorithm, term clustering,

* 연세대학교 문헌정보학과 대학원 (suy2onkim@hanmail.net)
** 연세대학교 문헌정보학과 교수 (ymchung@yonsei.ac.kr)

1. 서론

정보 과부하(information overload) 문제가 점점 심각해지면서 다양한 정보원으로부터 양질의 정보를 선택해야 하는 이용자의 역할이 커지고 있지만, 이용자는 자신이 원하는 정보를 적절한 질의로 잘 표현하지 못한다는 지적이 있어 왔다. Anick and Vaithyanthan(1997)은 이용자들의 정보요구는 다양한 요소들의 영향을 받기 때문에 질의에 잘 표현되지 못할 수도 있으며, 더욱이 정보요구 자체가 모호할 경우에는 그것을 질의어로 표현하는 데 더 큰 어려움을 겪는다고 주장하였다. 이런 문제를 해결하기 위한 질의확장(query expansion)은 효과적인 정보검색을 가능하게 할 뿐만 아니라 성공적인 검색을 위해 적절한 질의 표현을 만들어야 하는 이용자들의 부담을 덜어줄 수 있다.

초기 질의어와 연관된 용어를 선정하여 새로운 질의에 추가하는 자동 질의확장에 대한 연구는 1970년대에 들어 본격적으로 시작되었으며 최근에 더욱 활발해지고 있다(Rocchio 1971; Ide 1971; Qui and Frei 1993; Buckley, Salton, and Allan 1994; Xu and Croft 1996; Kim and Choi 1999; Chung and Lee 2004). 특히 전체 문헌집단으로부터 질의어와 연관된 용어를 통계적인 기준에 의해 선정하는 전역적(global) 방법은 시스템에 의해 자동적으로 처리될 수 있으므로 이용자의 부담이 없는 매우 효과적인 질의확장 기법이라 할 수 있다.

1990년대 후반에 들어서는 대량의 데이터로부터 상호 관련 있는 패턴이나 연관규칙(association rule)을 발견하기 위한 데이터 마이닝(data mining) 기법을 응용한 텍스트 마이닝(text mining)에 대한 연구가 수행되기 시작하였다. 데이터베이스 내 지식탐사(knowledge discovery in database) 과정에서 핵심적인 역할을 담당하는 데이터 마이닝은 대량의 데이터로부터 패턴인식, 통계적 기법, 인공지능 기법 등을 이용하여 숨겨져 있는 데이터 간의 상호 관련성 및 유용한 정보를 추출하는 단계이다(Michael and Gorden 1997).

텍스트 마이닝은 구조화된 데이터베이스를 대상으로 하는 데이터 마이닝과는 달리 구조화되지 않은 대규모의 텍스트 집단으로부터 새로운 지식을 발견하는 과정을 의미한다(정영미 2005). 다시 말해 텍스트 마이닝은 비구조화된 텍스트로부터 흥미 있고 유용한 패턴을 찾아내는 과정으로 볼 수 있으며, 정보추출, 정보검색, 자연언어 처리, 텍스트 요약, 자동분류 등에서 사용되는 기법들을 결합하여 사용한다(Tribula 1999). 텍스트 마이닝 적용의 핵심은 대량의 텍스트로부터 이전에 알려지지 않은 숨겨진 지식을 찾아내는 것이다.

예를 들어 과거에는 주로 유사계수를 이용한 용어 클러스터링을 통해 텍스트로부터 상호 연관된 용어들의 집합을 생성하였으나, 데이터 마이닝 기법인 연관규칙 마이닝

을 텍스트 처리에 응용함으로써 문헌 텍스트로부터 연관용어 집합을 생성하고, 이를 통해 유용한 지식 정보를 효과적으로 얻으려는 시도를 하게 된 것이다(Fayyad et al. 1996). 질의에 추가할 적절한 용어집합을 찾기 위해 연관규칙을 사용한 연구로 Rungsawang 등(1999)과 Wei 등(2000)의 연구가 있다. Rungsawang 등(1999)은 TREC의 하부컬렉션인 FT(Financial Times)를 이용하여 TREC7의 adhoc 질의를 가지고 Apriori 알고리즘을 적용해서 질의에 추가할 용어를 선정하였고, Wei 등(2000)은 자동적인 전역적 질의확장을 위해서 TREC4의 AP90 컬렉션을 이용하여 연관용어 마이닝 실험을 수행하였다. 이 연구에서는 이전의 용어 공기빈도에 기반한 연관용어 선정에 비해 연관용어 마이닝 방법은 신뢰도 계산까지 할 수 있기 때문에 연관 규칙을 이용해서 연관용어를 선정하는 방법이 좀 더 유용하다고 평가했다.

본 연구의 목적은 전역적 질의확장에서 초기 질의어와 의미적으로 가장 많이 연관된 용어를 추가 질의어로 사용할 수 있도록 하기 위해 최적의 성능을 보이는 연관용어 선정 기법을 찾아내는 데 있다. 연관용어 선정 기법으로는 연관규칙 마이닝 기법인 Apriori 알고리즘과 용어의 공기빈도를 기반으로 한 용어 클러스터링 기법을 사용하였다. 용어 클러스터링 기법은 질의확장을 위해 오래 전부터 적용되어 왔으며, 기본적으로 용어의 공기빈도에 기반하여 구축한 연관행렬(association matrix)을 사용한다(Baeza-Yates and Reibeiro-Neto 1999). 이때 연관행렬의 각 행은 특정한 질의어에 대한 연관 클러스터(association cluster)를 형성하며, 이 연관 클러스터로부터 추가 질의어 후보가 선정된다.

용어 간 연관성 척도로는 Apriori 알고리즘에서는 지지도(support) 및 신뢰도(confidence)와 향상도(lift)를 사용하며, 용어 클러스터링 기법에서는 고빈도 용어 선호 특성을 보이는 GSS 계수(Galavoti, Sebastiani & Simi 2000), 자카드계수(Jaccard coefficient), 코사인계수(cosine coefficient), 소칼 & 스니스 5(Sokal & Sneath 5)와 저빈도 용어를 선호하는 특성을 지닌 상호정보량(mutual information)을 사용하였다.

2. 연관용어 선정 기법

2.1 연관규칙 마이닝과 Apriori 알고리즘

연관규칙 마이닝은 대규모의 데이터 항목집합 사이에서 유용한 연관관계 및 상관관계를 찾는 방법이다. 연관규칙 마이닝의 가장 전형적인 활용 중 하나는 장바구니 분석(market basket analysis)이다. 이 분석기법은 고객들의 장바구니에서 서로 다른 품목들 사이의 연관관계를 발견함으로써 고객의 구매습관과 구매요구를 분석하는 것

이다(박우창 외 2003).

데이터 마이닝의 가장 일반적인 방법인 연관규칙을 텍스트 마이닝에 적용하기 위해서는 용어에 대한 정의를 새롭게 내릴 필요가 있다. 연관관계(association relationship)는 항목들 사이에 존재하는 유사성 또는 패턴을 의미한다. 데이터 마이닝에서의 연관규칙 기본개념은 데이터베이스에 n개의 트랜잭션 집합이 존재하고 이 트랜잭션에 포함된 모든 항목집합(itemset)을 I 라고 표현했을 경우 공집합이 아닌 항목집합 X, Y에 대해서 $X \subset I$, $Y \subset I$이고, $X \cap Y = \varnothing$ 일 경우 "조건 X라는 사건이 발생했을 때, 결과 Y라는 사건이 발생한다"는 것을 말하며 "X→Y" 로 표현된다.

이를 문헌 텍스트에 적용할 경우 여러 문헌에서 {a}와 {b}의 두 용어가 동시에 출현한다면 {a}와 {b}는 서로 연관성 있는 용어라는 것을 알 수 있다. 트랜잭션이란 발생된 데이터를 저장하는 단위이며 여러 항목을 가질 수 있는 개념인데 이를 문헌집단에 적용하면 한 문헌이 하나의 트랜잭션이 될 수 있다. 또한 이러한 개념을 단일 문헌에 적용한다면, 한 단락이나 한 문장이 하나의 트랜잭션이 될 수도 있다. n개의 트랜잭션들로 이루어진 데이터베이스(D)가 존재할 때 각 트랜잭션에는 고유한 트랜잭션 번호(TID: transaction id)가 부여된다(Agrawal and Srikant 1994). 문헌집단에 적용할 경우 각 트랜잭션을 구분해 주는 트랜잭션 id는 문헌번호(Doc id)와 같은 개념으로 볼 수 있다.

연관규칙이 성립되기 위해 필요한 조건은 규칙의 조건부와 결과부 모두 빈발해야 한다는 것이다. 주어진 데이터베이스에서 연관규칙을 찾는 것은 사용자가 정의한 최소지지도와 최소신뢰도 이상의 값을 갖는 항목집합을 찾는다는 의미이므로, 연관규칙을 찾는 과정은 기본적으로 빈발 항목집합을 찾는 단계와 연관규칙을 생성하는 두 단계로 구성된다(Agrawal 1993).

빈발 항목집합이란 후보 항목집합 중 최소지지도 및 최소신뢰도 임계치 이상의 값을 가진 항목집합으로서 연관성 마이닝 기법에서 생성되는 연관규칙의 단위가 된다. 항목집합이란 개별 트랜잭션에 포함된 단일항목 또는 복수항목집합이며, 후보 항목집합이란 개별 항목집합의 결합을 통해 생성된 항목집합으로 동시에 출현한 항목을 포함하는 집합이다. 항목집합의 발생빈도(occurrence frequency)는 항목집합을 포함하는 트랜잭션의 수로 정의하는데 이를 간단하게 표현하면 빈도수라고도 하며, 이는 지지도 개수 또는 항목집합의 개수(count of the itemset)에 해당한다.

문헌에서의 항목집합은 문헌이 포함하고 있는 용어들이 되고, 후보 항목집합은 동시에 출현한 용어를 포함하는 용어집합으로 정의내릴 수 있다. k-용어집합이란 k개의 용어로 이루어진 집합을 말하며, {데이터베이스, 관리}와 같은 용어집합은 "2-용어집합"이 된다. 항목집합의 발생빈도는 문헌집단에 적용할 경우 용어집합이 동시에 발생하는 트랜잭션의 수가 되므로 전체 문헌집단에서 용어가 동시에 출현한 문헌 수(문헌

빈도)가 된다. 빈발 항목집합은 후보 용어집합에서 최소지지도 및 최소신뢰도 이상의 값을 가진 용어집합(연관용어 집합)이 된다.

Agrawal and Srikant(1994)는 후보 항목집합을 생성하고, 발생 빈도를 계산한 후 사용자가 정의한 최소지지도를 가진 빈발 항목집합을 결정하는 Apriori 알고리즘과 AprioriTid 알고리즘을 제안하였다. 본 연구에서 사용한 Apriori 알고리즘은 각 패스에서 빈발 항목집합들의 후보 항목집합을 구성한 후에 각 후보 항목집합의 발생 빈도를 계산하고, 사용자가 정의한 최소지지도를 기준으로 하여 빈발 항목집합들을 결정한다. 다음 단계에서는 이들 빈발 항목집합들로부터 최소신뢰도 임계치를 만족하는 연관규칙을 모두 찾는다. Apriori 알고리즘에서 사용하는 중요한 법칙은 빈도수가 높은 항목집합의 모든 부분집합은 모두 빈도수가 높다는 사실이다. 어떤 집합이 주어졌을 때 새로운 항목을 더해주면 지지도는 절대로 전보다 증가할 수 없다. 예를 들어, 100건의 문헌집단 내에 {정보, 검색}이란 용어집합의 지지도가 32% (32건의 문헌에서 '정보'와 '검색'이 동시출현)였다면, {정보, 검색, 시스템}이란 용어집합의 지지도는 32% 보다 클 수 없다. 즉 '정보', '검색', '시스템'이 동시에 출현한 문헌 수는 절대 32건을 넘을 수 없음을 의미한다.

2.2 연관규칙 마이닝의 연관성 척도

연관규칙 마이닝의 연관성 척도로는 지지도, 신뢰도, 향상도를 사용한다. 연관규칙 X→Y에서 지지도란 전체 트랜잭션에 대한 X와 Y를 만족하는 트랜잭션의 비율을 의미한다. 지지도가 높을수록 데이터베이스 내에 해당 트랜잭션이 더 많이 포함되어 있다는 것을 나타낸다. 이러한 지지도는 규칙의 통계적 중요성으로 인식될 수 있으며 빈번하게 발생되는 패턴 또는 규칙의 빈도를 의미하므로 그 패턴이나 규칙의 유용성이 증대되려면 지지도가 커야 한다는 것을 알 수 있다. 지지도는 지지도 개수(support count)로 간단하게 표현하기도 하는데 지지도 개수란 항목이 출현한 트랜잭션의 수를 의미한다.

신뢰도는 X를 포함하는 트랜잭션에 대한 Y를 포함하는 트랜잭션의 비율을 의미한다. 이러한 신뢰도는 규칙 X→Y에 대한 정확도를 측정할 수 있는 지표가 되며, 높은 신뢰도는 정확한 예측을 가능하게 해준다. 지지도가 낮아도 신뢰도가 높을 경우 유용한 연관규칙일 가능성이 높다. 결론적으로 어떤 규칙의 신뢰도는 조건부에 대해 결과부가 얼마나 자주 적용될 수 있는지를 나타내고, 반면에 지지도는 그 규칙 자체가 얼마나 믿을만한지를 나타낸다고 할 수 있다.

향상도는 X, Y의 상관관계를 나타내는 척도로서, 분자는 X, Y가 동시에 트랜잭션에 존재할 확률을 의미하며 분모는 각 X, Y가 독립적으로 트랜잭션에 존재할 확률을 의

미한다. 향상도 값이 1이면 X와 Y가 서로 독립적임을 의미하고, 1보다 큰 경우에는 X와 Y가 서로 양의 상관관계임을 나타내는 것이며, 1보다 작을 경우에는 X와 Y가 서로 음의 상관관계임을 의미한다.

2.3 용어 클러스터링의 연관성 척도

본 연구에서 용어 클러스터링에 의한 연관용어 선정 실험에서는 용어 간 연관성 척도로 많이 사용되고 있는 자카드계수, 코사인계수, 소칼 & 스니스 5와 자동분류의 자질선정 기준으로 사용되어 좋은 성능을 보인 GSS 계수를 선정하였다(Golavotti et al. 2000). 자카드계수, 코사인계수, GSS 계수, 소칼 & 스니스 5는 연관성 척도의 빈도수준 선호경향에 대한 연구에서 고빈도 용어를 선호하는 유사계수로 나타났으며(이재윤 2004), 고빈도 용어를 선호하는 Apriori 알고리즘과 비교분석하기에 적합할 것으로 판단되어 선정하였다. 또 저빈도 용어를 선호하는 상호정보량도 용어 클러스터링을 위한 연관성 척도로 사용하였다.

실험에 사용한 유사계수 공식은 <표 1>과 같다. <표 1>에서 a는 용어 X와 용어 Y가 함께 출현한 문헌 수를, b는 용어 X는 출현하지 않고 용어 Y만 출현한 문헌 수를, c는 용어 X는 출현하고 용어 Y는 출현하지 않은 문헌 수를, d는 두 용어 모두 출현하지 않은 문헌 수를 각각 나타낸다.

<표 1> 용어 클러스터링의 유사계수 공식

명칭	약호	공식
자카드계수	JAC	$\frac{a}{a+b+c}$
코사인계수	COS	$\frac{a}{\sqrt{(a+b)(a+c)}}$
GSS 계수	GSS	$\frac{ad-bc}{N^2}$
소칼 & 스니스 5	SS5	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
상호정보량	MI	$\log_2\frac{Na}{\sqrt{(a+b)(a+c)}}$

3. 연관용어 선정 실험

3.1 실험데이터

실험문헌집단으로 KTSET 1.0을 이용하였으며, 각 문헌 레코드의 한글 제목과 초록 필드로부터 한글 형태소 분석기인 HAM을 사용하여 용어를 추출하였다. 제목과 초록에 출현한 용어들은 주제적인 성격을 가지고 있기 때문에 연관용어 집합에 주제적 특성이 적은 일반적인 용어들이 포함되는 문제를 어느 정도 해결할 수 있다.

실험문헌집단에서 제공하는 30개의 질의 중 정보학과 관련된 20개의 질의어 집합(<표 2>)과 7개의 질의 집합(<표 3>)을 선택하여 연관용어 선정 실험을 수행하였으며, 연관용어 선정 범위는 전체 문헌집단 중 5건 이상의 문헌에 출현한 용어들로 제한하였다. 그 이유는 예비실험에서 문헌빈도가 5 미만인 용어들까지 연관용어의 선정 범위로 포함했을 때 주제적으로 연관이 적은 용어집합이 많이 생성되었기 때문이다.

<표 2> 연관용어 선정 실험을 위한 질의어 집합

질의어 번호	질의어	문헌빈도	질의어 번호	질의어	문헌빈도
1	데이타베이스	122	11	marc	6
2	멀티미디어	51	12	시스템	474
3	정보검색	20	13	지능형	11
4	한국어	35	14	도서관	35
5	검색	107	15	전문가	31
6	분석	250	16	검색효율	6
7	언어학적	3	17	이론	34
8	자동색인	7	18	정보시스템	4
9	효율성	34	19	집합	47
10	kormarc	7	20	퍼지	25

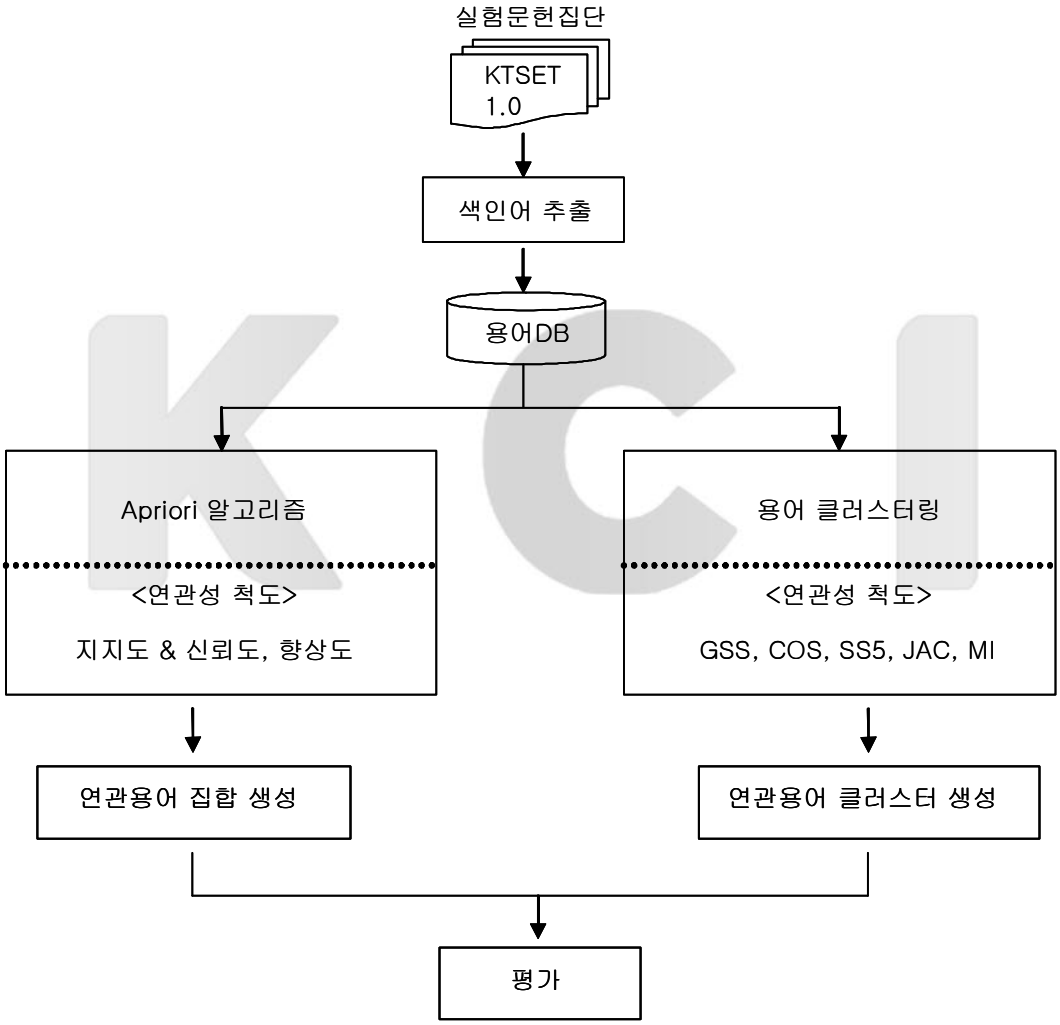
<표 3> 연관용어 선정 실험을 위한 질의 집합

질의 번호	질의어	질의어 수	적합문헌 수	적합문헌 색인어수
Q2	멀티미디어, 데이타베이스	2	25	490
Q22	한국어, 정보검색	2	18	587
Q26	언어학적, 분석, 자동색인, 검색, 효율성	5	5	165
Q27	kormarc, marc	2	5	193
Q28	지능형, 정보검색, 시스템	3	8	196
Q29	전문가, 시스템, 도서관	3	4	96

Q30	퍼지, 집합, 이론, 정보시스템, 검색효율	5	2	57

3.2 실험개요

<그림 1>은 텍스트 마이닝 기법을 이용해서 초기 질의어의 연관용어를 선정하기 위한 전체적인 실험개요이다.



<그림 1> 연관용어 선정 실험 개요

본실험에 앞서 수행한 예비실험에서 지지도 및 신뢰도 임계치에 따라 생성된 연관

용어 집합을 살펴본 결과, Apriori 알고리즘을 이용한 연관용어를 선정 시 지지도 및 신뢰도의 임계치를 질의어에 일률적으로 적용할 경우, 질의어에 대한 연관용어가 한 개도 생성되지 않는 경우들이 있었다. 그 이유는 Apriori 알고리즘의 지지도 및 신뢰도 척도는 질의어와 연관용어가 전체문헌집단에 출현한 빈도에 영향을 받기 때문에 주어진 질의어보다 지지도 임계치가 낮을 경우 연관용어 집합이 생성되지 않는 것이다. 예를 들면 지지도의 임계치가 8로 정해진 경우(질의어와 연관용어의 공기빈도가 8 이상) 초기 질의어의 문헌빈도가 3이면 연관용어 집합이 생성되지 못한다. 따라서 이 연구에서는 각 연관성 척도의 실험결과를 비교, 평가하기 위해서 연관용어 집합에 연관용어가 1개 포함될 경우 연관용어 집합이 10개 이상 생성될 수 있도록 질의어에 따라 임계치를 달리하였다.

연관용어 집합의 크기를 한 항목(용어)씩 늘려가는 Apriori 알고리즘과 연관용어 선정 성능을 비교하기 위해서 유사계수를 이용해 생성한 연관용어 클러스터도 한 클러스터에 포함되는 연관용어 수를 한 개씩 늘려가면서 생성하도록 하였다. 이 연구에서는 질의어와 연관된 용어의 수를 1개부터 3개까지 늘려가면서 실험하였다. 즉, 연관용어 집합의 크기를 2-연관용어 집합(L2)부터 4-연관용어 집합(L4)까지 생성하고, 연관용어 클러스터도 한 클러스터에 초기 질의어와 연관용어 1개를 선정한 2-연관용어 클러스터(C2)부터 연관용어를 3개까지 선정하는 4-연관용어 클러스터(C4)까지 생성하였다. 예를 들면, 질의어 Q_1에 대해 연관용어를 1개 선정할 경우 연관용어 집합은 $\{Q_1, t_1\}$, $\{Q_1, t_2\}$, …, $\{Q_1, t_n\}$ 와 같이 생성되며, 3개의 연관용어를 선정할 경우에는 $\{Q_1, t_1, t_2, t_3\}$, $\{Q_1, t_1, t_2, t_n\}$, …, $\{Q_1, t_x, t_y, t_z\}$ 와 같이 생성된다.

3.3 실험결과 평가방법

각 텍스트 마이닝 기법을 이용하여 20개의 질의어에 대한 연관용어 집합을 각각 생성한 후 성능 평가를 위해 상위 10위까지의 연관용어 집합을 이용하였다. 상위 n개의 연관용어 집합에 대해 적합성 판정을 할 경우 유사도 값이 같음에도 불구하고 평가에서 제외되는 용어들이 생기기 때문에 같은 유사도 값을 갖는 연관용어에는 동순위를 부여하였다.

실험 과정에서 연관용어의 수를 한 개씩 늘려가면서 연관성을 측정하게 되는데 용어 X, Y, Z의 연관성을 측정할 경우 이론적으로는 {(X, Y), Z}, {(X, Z), Y}, {(Y, Z), X}와 같이 세 가지 경우의 수가 발생하지만, 특정 질의어 X에 대한 연관용어 클러스터를 생성할 경우에는 각각 {(X, Y), Z}, {(X, Z), Y} 두 경우의 유사도 값이 측정된다. 이때, 측정된 유사도 값은 다르지만, 각각의 경우 유사도 값이 상위 10위 내에 해당되어 모두 평가대상이 될 수 있다는 문제가 있다. 이런 문제를 해결하기 위해서, 특

정 질의어에 대해 생성된 연관용어 집합이 모두 동일한 연관용어들을 포함하고 있을 경우 유사도 값이 최대가 되는 경우만을 연관용어 집합에 포함시켰다.

각 텍스트 마이닝 기법의 연관용어 선정 성능을 평가하는 방법으로 연관용어 정확률과 연관용어 일치율을 이용하였다. 연관용어 정확률은 20개의 질의어에 대해 텍스트 마이닝 기법의 각 연관성 척도가 선정한 10위 내의 연관용어가 주제전문가에 의해 적합하다고 판정된 연관용어 집합에 포함될 비율을 나타낸다. 주제전문가로서 10명의 문헌정보학과 대학원생이 연관용어 적합성 판정을 했으며, 10명 중 8명 이상이 각 질의어에 대해 연관된 용어라고 판정한 용어들을 적합한 연관용어로 보았다.

연관용어 일치율은 실험문헌집단에서 각 질의에 대해 적합문헌으로 판정된 문헌들의 색인어와 텍스트 마이닝 기법의 각 연관성 척도가 선정한 10위 내 연관용어가 얼마나 일치하는지를 알아본 것이다. 실험문헌집단에서는 각 질의에 대해서만 적합문헌이 제시되어 있기 때문에 7개의 질의를 대상으로 연관용어 일치율을 구하였다. 연관용어 정확률 공식과 연관용어 일치율 공식은 각각 다음과 같다.

$$\text{연관용어 정확률} = \frac{\text{전문가에 의해 판정된 연관용어 수}}{\text{10위 내 연관용어 수}}$$

$$\text{연관용어 일치율} = \frac{\text{적합문헌 색인어와 일치하는 연관용어 수}}{\text{10위 내 연관용어 수}}$$

각 질의어 및 질의에 대한 정확률과 일치율을 구한 후에 각 연관성 척도의 성능을 평가하기 위해서 흔히 사용되는 매크로 평가방법을 적용하여 연관용어 평균정확률과 평균일치율을 구하였다.

마지막으로 각 연관성 척도가 선정한 연관용어들이 서로 얼마나 유사한지 알아보기 위해 연관성 척도 간 연관용어 일치율을 알아보았다. 연관성 척도 A를 기준으로 연관성 척도 A와 B가 선정한 연관용어 일치율을 구할 경우 공식은 다음과 같다.

$$\text{연관성 척도 일치율 (A}\Rightarrow\text{B)} = \frac{\text{연관성 척도 } A\text{와 } B\text{의 중복 연관용어 수}}{\text{연관성 척도 } A\text{의 연관용어 수}}$$

4. 실험결과 분석

4.1 전문가에 의한 연관용어 적합성 평가

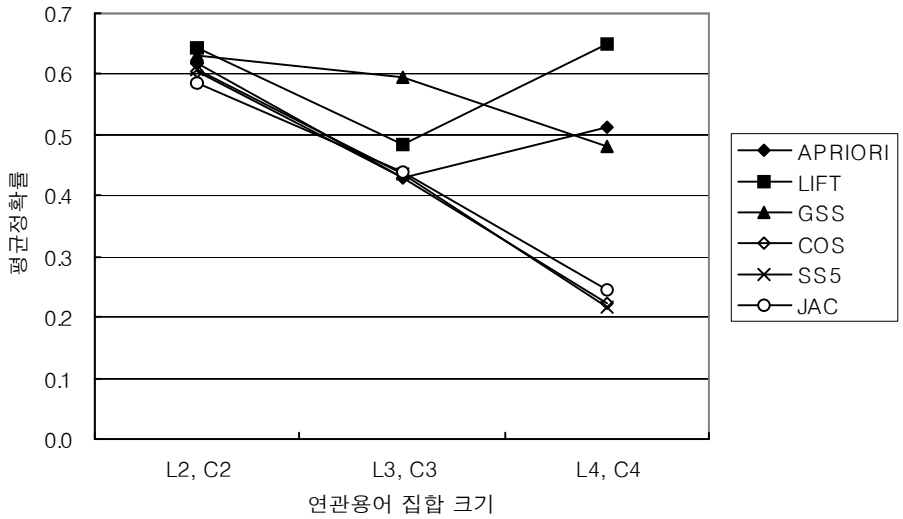
<표 4>와 <그림 2>는 각 연관성 척도가 20개의 질의어에 대해 선정한 연관용어 집합들 중 상위 10위에 속하는 연관용어 집합을 대상으로 연관용어 평균정확률을 구한 결과를 보여 준다. 분석 결과 첫째, Apriori 알고리즘의 척도인 지지도 및 신뢰도(APRIORI)와 향상도(LIFT)를 이용해서 연관용어 집합을 생성할 경우 집합의 크기(L2, L3, L4)에 상관없이 향상도 척도가 지지도 및 신뢰도 척도에 비해 더 높은 평균정확률을 보였다. 둘째, GSS 계수(GSS)가 모든 집합크기에서 비교적 높은 평균정확률을 보였으며, 셋째, 유사계수에 의해 연관용어 집합을 생성했을 경우에는 집합의 크기가 커질수록 GSS 계수(GSS), 코사인계수(COS), 소칼 & 스니스 5(SS5), 자카드계수(JAC) 모두 평균정확률이 점차 낮아지는 경향을 나타냈다. 넷째, 연관용어 1개만을 포함하는 L2/C2 집합에서 지지도 및 신뢰도(APRIORI) 척도가 0.617, 향상도(LIFT) 척도가 0.642, GSS 계수(GSS)가 0.631, 코사인계수(COS)가 0.606, 소칼 & 스니스 5가 0.607, 자카드계수(JAC)가 0.584의 평균정확률을 보여 연관성 척도별 차이가 그다지 크지 않은 것으로 나타났다. 또한 전체적으로 보아 GSS 계수(GSS)나 향상도(LIFT), 지지도 및 신뢰도(APRIORI)에 의해 연관용어를 선정할 경우가 코사인계수(COS)나 자카드계수(JAC), 소칼 & 스니스 5(SS5)를 사용할 경우에 비해 높은 평균정확률을 보였다.

<표 4> 연관용어 평균정확률

연관성 척도	평균정확률		
	L2/C2	L3/C3	L4/C4
APRIORI	0.617	0.428	0.511
LIFT	0.642	0.485	0.650
GSS	0.631	0.596	0.480
COS	0.606	0.428	0.222
SS5	0.607	0.436	0.216
JAC	0.584	0.440	0.244

4.2 적합문헌을 이용한 연관용어 적합성 평가

<표 5>는 적합문헌을 이용한 적합성 평가 결과를 보여 준다. 연관용어를 1개 선정할 경우 적합문헌에 의한 평가에서는 GSS 계수(GSS)가 모든 연관성 척도 중에 가장 높은 연관용어 일치율을 보였고, 다음으로 Apriori 알고리즘의 척도인 향상도(LIFT)와 지지도 및 신뢰도(APRIORI)가 높은 성능을 나타냈다. 유사계수 중에는 자카드계수(JAC)가 다른 유사계수들(COS, SS5, MI) 보다 비교적 좋은 성능을 보이는 것으로 나타났다. 특이한 점은 질의확장에서 좋은 성능을 보인다고 보고된 상호정보량(MI)의



<그림 2> 연관성 척도별 평균정확률

성능이 이 실험에서는 낮게 나타났다는 것이다. 그 이유는 실험에서 상호정보량에 의해 선정된 연관용어를 10위까지 선택할 경우 다른 연관성 척도보다 월등히 많은 용어가 포함되어서 연관용어 일치율이 다른 연관성 척도에 비해 상대적으로 낮게 평가될 수밖에 없기 때문이다. 상호정보량을 제외한 다른 연관성 척도에 의해 생성된 연관용어 수는 각 질의어에 대해 평균 10개 정도이다. 그러나 상호정보량을 이용할 경우에는 각 질의어에 대해 20개의 연관용어가 생성된다. 특히, Q26에서 133개, Q29에서 88개, Q30에서 117개 등 많은 수의 연관용어가 10위 내 연관용어로 선정되었는데, 이 가운데 적합문헌 색인어와 중복된 고유용어 수는 Q26에서 35개, Q29에서 11개, Q30에서 6개로서 연관용어 일치율은 각각 0.26, 0.13, 0.05로 낮게 나타났다.

<표 5> 연관용어 일치율

질의 번호	질의어 수	연관성 척도	연관용어 1개 선정 시		연관용어 2개 선정 시		연관용어 3개 선정 시	
			연관용어 일치율	질의별 평균	연관용어 일치율	질의별 평균	연관용어 일치율	질의별 평균
Q2	2	APRIORI	1	0.93	1	0.89	1	0.84
		LIFT	0.94		1		1	
		GSS	1		1		1	
		COS	0.95		0.92		0.84	

		SS5	0.89		0.83		0.81	
		JAC	1		0.92		0.84	
		MI	0.71		0.53		0.36	
Q22	2	APRIORI	0.90	0.75	0.79	0.68	0.78	0.64
		LIFT	0.65		0.75		0.71	
		GSS	0.91		0.89		0.87	
		COS	0.69		0.60		0.61	
		SS5	0.68		0.62		0.59	
		JAC	0.81		0.65		0.57	
		MI	0.59		0.46		0.38	
Q26	5	APRIORI	0.80	0.63	0.79	0.58	0.89	0.49
		LIFT	0.72		0.71		0.84	
		GSS	0.76		0.77		0.80	
		COS	0.61		0.55		0.28	
		SS5	0.57		0.58		0.29	
		JAC	0.67		0.45		0.18	
		MI	0.26		0.18		0.11	
Q27	2	APRIORI	1	0.98	1	0.96	1	0.91
		LIFT	1		1		1	
		GSS	1		1		1	
		COS	1		1		0.82	
		SS5	1		1		0.83	
		JAC	1		0.90		0.86	
		MI	0.83		0.79		0.86	
Q28	3	APRIORI	0.52	0.34	0.39	0.31	0.45	0.28
		LIFT	0.25		0.31		0.42	
		GSS	0.40		0.42		0.34	
		COS	0.33		0.31		0.24	
		SS5	0.31		0.26		0.25	
		JAC	0.38		0.33		0.21	
		MI	0.18		0.16		0.05	
Q29	3	APRIORI	0.55	0.45	0.62	0.45	0.71	0.4
		LIFT	0.42		0.51		0.75	
		GSS	0.61		0.62		0.62	
		COS	0.46		0.44		0.21	
		SS5	0.46		0.39		0.20	
		JAC	0.53		0.45		0.21	
		MI	0.13		0.11		0.08	
Q30	5	APRIORI	0.25	0.18	0.21	0.13	0.30	0.14
		LIFT	0.28		0.27		0.32	
		GSS	0.36		0.30		0.23	
		COS	0.13		0.06		0.04	
		SS5	0.12		0.04		0.04	
		JAC	0.09		0.05		0.04	
		MI	0.05		0.02		0.03	

연관용어를 2개 선정할 경우에도 GSS 계수(GSS)가 제일 높은 성능을 보이고 있으며, 향상도(LIFT)와 지지도 및 신뢰도(APRIORI)가 비교적 높은 성능을 나타낸다. 상호정보량(MI)은 연관용어를 2개 선정할 경우에도 다른 연관성 척도보다 많은 연관용어 집합을 생성하는 경향을 나타낸다. 특히, Q28, Q29, Q30의 경우에 다른 연관성 척도가 선정한 10위 내 고유용어 수에 비해서 상호정보량(MI)을 이용했을 경우 약 100개 정도 더 많은 연관용어들이 선정되었다. 용어 간 연관성을 측정하는 데 자주 사용되는 자카드계수(JAC)나 코사인계수(COS)는 연관용어 일치율이 GSS 계수(GSS)를 이용한 경우보다 좋지 않았으며, 코사인계수(COS), 자카드계수(JAC), 소칼 & 스니스 5(SS5)는 Q27을 제외하고 모두 비슷한 성능을 보였다.

연관용어를 3개까지 선정했을 경우에도 마찬가지로 GSS 계수(GSS), 향상도(LIFT), 지지도 및 신뢰도(APRIORI) 척도를 이용할 경우에 높은 연관용어 일치율을 보였다. 또한 코사인계수(COS), 자카드계수(JAC)와 소칼 & 스니스 5(SS5)는 서로 비슷한 수준의 연관용어 일치율을 보이고 있다.

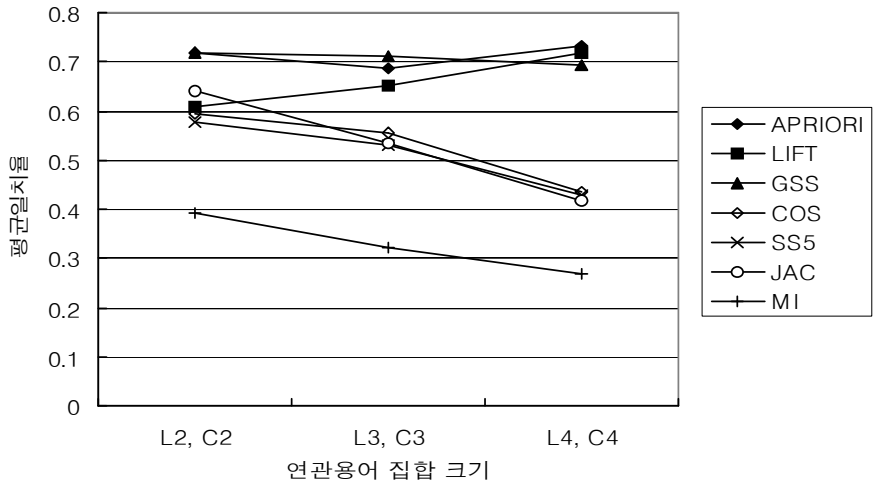
3개의 연관용어 선정 시에도 마찬가지로 Q27에 대한 연관용어 일치율이 가장 높았으며, Q30에 대한 연관용어 일치율이 가장 낮았다. 특이한 점은 질의어에 추가되는 연관용어의 수가 늘어날수록 Q27에 대한 연관용어 일치율 평균은 0.98, 0.96, 0.91로 낮아졌으나 Q30에 대한 연관용어 일치율 평균은 0.18, 0.13으로 낮아지다가 연관용어가 3개 선정될 경우에는 다시 0.14로 조금 높아졌다는 것이다. 그 이유는 Q30의 경우 유사계수를 이용하여 연관용어를 선정했을 경우에는 연관용어 일치율이 Q27과 마찬가지로 점차 낮아졌지만 향상도(LIFT)와 지지도 및 신뢰도(APRIORI)를 이용해 연관용어를 선정했을 경우의 연관용어 일치율이 향상도의 경우 L2에서 0.28, L3에서 0.27로 낮아지다가 L4에서 다시 0.32으로 높아졌고, 지지도 및 신뢰도 척도를 이용했을 경우도 마찬가지로 L2에서 0.25, L3에서 0.21로 낮아졌다가 L4에서 0.30으로 높아졌기 때문으로 보인다.

<표 6>과 <그림 3>은 각 연관성 척도를 이용해 7개의 질의에 대해 선정한 연관용어들 중 상위 10위에 해당하는 연관용어 집합을 대상으로 연관용어 평균일치율을 구한 결과이다. 분석 결과 첫째, GSS 계수(GSS)와 지지도 및 신뢰도(APRIORI)를 이용하여 연관용어 집합을 생성했을 때가 모든 집합크기에서 비교적 높은 평균일치율을 나타냈다. 둘째, 유사계수(GSS, COS, SS5, JAC, MI)들을 이용하여 연관용어 집합을 생성했을 경우 집합의 크기가 커짐에 따라서 연관용어와 적합문헌 색인어와의 평균일치율이 점차 낮아졌다. 그러나 향상도(LITF)를 이용했을 경우는 집합의 크기가 커짐에 따라서 연관용어 평균일치율이 점차 높아지는 경향을 보였다. 셋째, 코사인계수

(COS)와 자카드계수(JAC), 소칼 & 스니스 5(SS5)를 이용했을 경우 동일한 집합의 크기에서 세 가지 연관성 척도가 비슷한 수준의 평균일치율을 나타냄이 확인되었다. 넷째, 상호정보량(MI)에 의해 연관용어 집합을 생성했을 경우 가장 낮은 평균일치율이 산출되었다.

<표 6> 연관용어 평균일치율

연관성 척도	평균일치율		
	L2/C2	L3/C3	L4/C4
APRIORI	0.717	0.685	0.733
LIFT	0.609	0.65	0.72
GSS	0.719	0.713	0.695
COS	0.596	0.554	0.436
SS5	0.576	0.532	0.428
JAC	0.64	0.535	0.417
MI	0.393	0.322	0.269



<그림 3> 연관성 척도별 평균일치율

4.3 연관성 척도 일치율 분석

<표 7>과 <그림 4>는 선정된 연관용어들의 연관성 척도 간 일치율을 나타낸 것이다. 분석 결과 첫째, GSS 계수(GSS)가 다른 유사계수들에 비해 Apriori 알고리즘의 지지도 및 신뢰도(APRIORI) 척도와 향상도(LIFT) 척도가 선정한 연관용어와 동일한 연관용어를 가장 많이 선정하였다. 둘째, 코사인계수와 소칼 & 스니스 5와의 일치율이 각각 0.969, 0.949로서 이 두 연관성 척도가 선정한 연관용어들이 가장 많이 일치했음을 알 수 있으며, 셋째, 향상도(LIFT)는 연관용어를 선정하는 데 있어 고빈도어를 선호하는 유사계수들과 비슷한 경향을 보이는 것으로 나타났다.

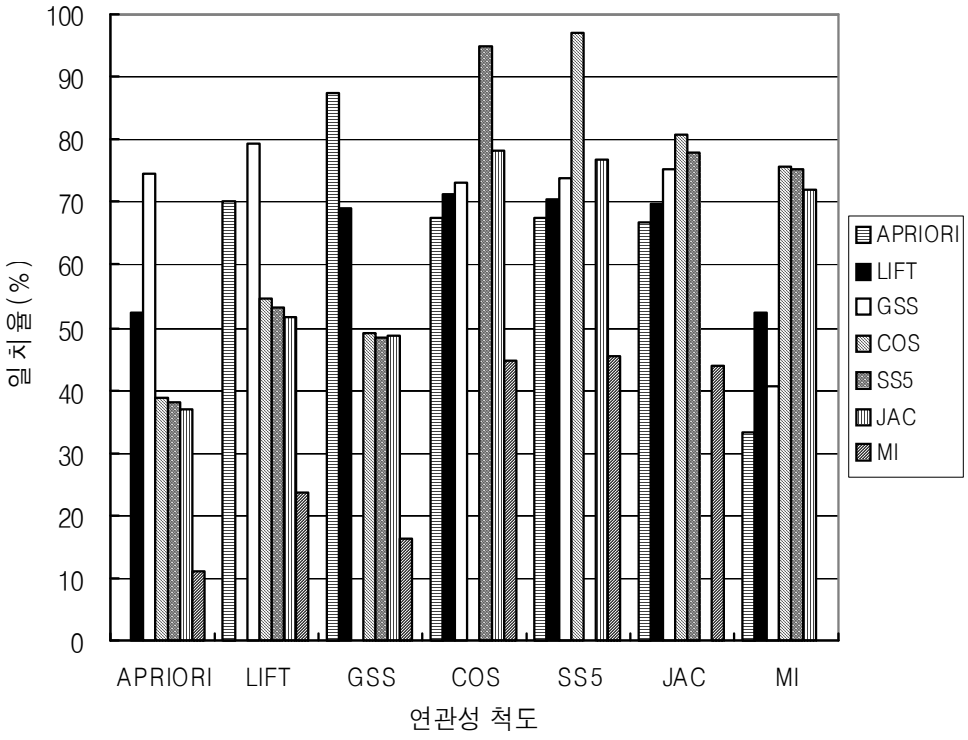
<표 7> 연관성 척도 일치율

	APRIORI	GSS	LIFT	COS	SS5	JAC	MI
APRIORI	1.000	0.746	0.523	0.387	0.379	0.368	0.112
GSS	0.874	1.000	0.691	0.490	0.485	0.488	0.161
LIFT	0.703	0.792	1.000	0.546	0.530	0.517	0.237
COS	0.676	0.731	0.711	1.000	0.949	0.781	0.447
SS5	0.676	0.738	0.705	0.969	1.000	0.766	0.453
JAC	0.667	0.754	0.698	0.809	0.778	1.000	0.441
MI	0.333	0.408	0.523	0.758	0.753	0.721	1.000

5. 결론

이 연구에서는 텍스트 마이닝 기법을 이용하여 연관용어 선정 실험을 수행하였다. 전문가에 의한 연관용어 적합성 판정에서 텍스트 마이닝 기법의 각 연관성 척도에 의해 생성된 크기가 다른 3개의 연관용어 집합의 성능을 평가한 결과, 연관용어 평균정확률은 Apriori 알고리즘의 향상도 척도에 의해 선정된 용어들이 0.592로 가장 높았으며, 다음으로 GSS 계수가 0.569로 높은 성능을 보였다. 집합크기별로 살펴보면 향상도 척도를 제외하고는 연관용어 집합에 포함된 연관용어의 수가 1개일 때 평균정확률이 가장 높았다.

적합문헌 색인어와 선정된 연관용어의 평균일치율은 크기가 다른 3개의 연관용어 집합의 성능평가 결과 Apriori 알고리즘의 지지도 및 신뢰도 척도가 0.711, GSS 계수가 0.709로 다른 척도에 비해 좋은 성능을 나타냈다. 집합 크기별로 살펴보면 Apriori 알고리즘의 척도인 지지도 및 신뢰도와 향상도는 연관용어 집합에 연관용어 3개가 포함되었을 때 평균일치율이 가장 높았다. 그러나 유사계수를 이용하여 연관용어 클러스터를 생성했을 때는 연관용어 집합에 연관용어가 1개 포함되었을 경우 평균일치율이 가장 높게 나타났다.



<그림 4> 연관성 척도 일치율

각 연관성 척도에 의해 생성된 상위 10위까지의 연관용어 집합에 포함된 연관용어들의 일치율을 살펴보면, 코사인계수와 소칼 & 스니스 5 계수가 선정한 용어들이 96.9%로 가장 높은 일치율을 보였으며, Apriori 알고리즘의 지지도 및 신뢰도 척도와 GSS 계수가 선정한 용어들도 87.4%로 비교적 높은 일치율을 나타냈다.

결론적으로 연관규칙 마이닝 기법인 Apriori 알고리즘의 지지도 및 신뢰도 척도와 향상도 척도에 의해 선정된 연관용어가 정확률과 일치율에서 비교적 좋은 성능을 보였으며, GSS 계수가 이와 비슷한 성능을 나타냈다. 코사인계수, 자카드계수, 소칼 & 스니스 5는 상대적으로 낮은 수준의 성능을 보였으며, 상호정보량은 가장 좋지 않은 성능을 보였다. 이 연구에서 Apriori 알고리즘은 대체로 좋은 성능을 보였지만 최소지지도와 최소신뢰도 임계치에 따라 생성된 연관용어 집합이 크게 달라지는 결과를 가져왔으며, 또한 적절한 임계치를 선정할 수 있는 일관된 기준이 없다는 것이 큰 문제점으로 나타났다.

이 연구는 클러스터링 기법에서 자주 이용되는 코사인계수, 자카드계수 등 모두 5가지의 연관성 척도와 연관규칙 마이닝 기법의 지지도 및 신뢰도와 향상도 척도를 이용하여 연관용어 선정 실험을 수행하여 그 결과를 비교, 평가하였다는 데 의의가 있다. 이 연구에서 다양한 연관성 척도에 의해 생성된 연관용어 집합은 질의확장이나 자동 시소러스 구축, 이용자가 제시한 질의어에 대한 연관용어 추천 등 정보검색 환경에서 다양하게 이용될 수 있을 것이다.

그러나 이 연구에서는 각 연관성 척도가 연관용어를 선정하는 성능만을 평가하였고, 선정된 연관용어를 추가 질의어로 사용한 질의확장은 수행하지 않았다. 실제로 질의확장에 관한 선행연구에서 좋은 검색 성능을 보였던 상호정보량은 이 연구에서는 가장 낮은 연관용어 선정 성능을 보였다. 기존 연구의 실험결과를 살펴보면 초기 질의어에 추가되는 용어의 수가 많아질수록 상호정보량이 좋은 성능을 나타내는 경향을 보였는데 이 연구에서는 다른 연관성 척도와의 비교를 위해 선정할 연관용어의 수를 제한하였기 때문에 다른 결과를 보였을 가능성이 있다. 따라서 앞으로의 연구에서는 연관성 척도에 의해 선정된 연관용어를 추가 질의어로 사용한 검색 실험을 수행함으로써 질의확장에서 있어서 각 연관성 척도의 효용성을 평가할 필요가 있다.

참 고 문 헌

박우창, 승현우, 용환승. 2003. 『데이터마이닝: 개념 및 기법』. 서울: 자유아카데미.

이재윤. 2004. 연관성 척도의 빈도수준 선호경향에 대한 연구. 『정보관리학회지』, 21(4): 281-294.

정영미. 2005. 『정보검색연구』. 서울: 구미무역 출판부.

Agrawal, R., Imielinski, T., and Swami, A. 1993. “Mining Association Rules between Sets of Items in Large Database." *Proceeding of the ACM SIGMOD International Conference on Management of Data, Washington, U.S.A.,* 207-216.

Agrawal, R., and Srikant, R. 1994. “Fast Algorithms for Mining Association Rules.” *Proceeding of the 20th International Conference on Very Large Databases, Santiago, Chile, September.*

Anick, P. G., and Vaithyanathan, S. 1997. “Exploiting Clustering and Phrases for Context-Based Information Retrieval.” *Proceeding of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 314-323.

Baeza-Yates, R. and Ribeiro-Neto, B. 1999. *Modern Information Retrieval.*

ACM Press.

Buckley, C., Salton, G., and Allan, J. 1994. "The Effect of Adding Relevance Information in a Relevance Feedback Environment." *Proceeding of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 292-300.

Chung, Y.M. and Lee, J.Y. 2004. Optimization of Some Factors Affecting the Performance of Query Expansion. *Information Processing and Management,* 40(6): 891-917.

Fayyad, U.M., Piatesky-shapiro, G., Smyth, P., and Uthurusamy, R. 1996. *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Galavoti, L., Sebastiani, F., and Simi, M. 2000. "Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization." *Proc. of ECDL-00. 4th European Conference on Research and Advanced Technology for Digital Libraries(Lisbon, Portugal, 2000):* 59-68.

Ide, E. 1971. "New Experiments in Relevance Feedback." In Salton, G., ed. *The SMART Retrieval System Experiments in Automatic Document Processing.* 337-354.

Kim, M.C. and Choi, K.S. 1999. "A Comparison of Collocation-Based Similarity Measures in Query Expansion." *Information Processing and Management,* 35(1): 19-30.

Lesk, M. E. 1969. Word-Word "Association in Document Retrieval Systems." *American Documentation,* 20(1): 27-38.

Michael, J., and Gorden, L. 1997. *Data Mining Techniques: For Marketing, Sales, and Customer Support.* John Wiley & Sons, Inc.

Peat, H, J., and Willett, P. 1991. "The Limitation of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems." *Journal of the American Society for Information Science,* 42(5): 378-383.

Qiu, Y., and Frei, H.P. 1993. "Concept Based Query Expansion." *Proceedings of the 16th Annual International ACM SIGIR conference on Research and Development in Information Retrieval.* 160-169.

Rocchio, J.J. 1971. "Relevance Feedback in Information Retrieval." In Salton, G., ed. *The SMART Retrieval System Experiments in Automatic Document Processing.* 313-323.

Rungsawang, A., Tangpong, A., Laohawee, P., Khampachua, T. 1999. Novel

Query Expansion Technique using Apriori Algorithm. <http://www.informatik.uni-trier.de/~ley/db/conf/trec/trec1999.html>

Tribula, W.J. 1999. “Text Mining.” *Annual Review of Information Science and Technology*, 34: 385-419.

Wei, J., Bressan S., and Ooi B.C. 2000. Mining Term Rules for Automatic Global Query Expansion: Methodology and Preliminary Results. *Web Information Systems Engineering, Proceeding of the First International Conference,* 1: 366-373.

Xu, J. and Croft, W.B. 1996. "Query Expansion using Local and Global Document Analysis." *Proceedings of the 19th Annual International ACM SIGIR conference on Research and Development in Information Retrieval.* 4-11.

KCI