논문 검색 결과의 효과적인 브라우징을 위한 단어 군집화 기반의 결과 내 군집화 기법

(A Search-Result Clustering Method based on Word Clustering for Effective Browsing of the Paper Retrieval Results)

배 경 만 † 황 재 원 †† 고 영 중 ††† 김 종 훈 †††
(Kyoungman Bae) (Jaewon Hwang) (Youngjoong Ko) (Jonghoon Kim)

요 약 검색 결과 내 군집화(search-result clustering)는 검색 엔진으로부터 검색된 결과 내에서 비슷한 문서를 자동으로 군집화하는 기법이다. 본 논문에서는 논문 검색 서비스에 전문화된 새로운 결과 내 군집화 기법을 제안한다. 제안하는 시스템은 '범주체계생성기(Category Hierarchy Generation System)'와 '논문군집기(Paper Clustering System)'로 구성되어있다. '범주체계생성기'는 KOSEF의 연구 범주 체계를 이용하여 분야 시소러스라 불리는 범주 체계를 생성하고, K-means 알고리즘을 이용한 단어 군집화 알고리즘을 사용하여 분야 시소러스의 키워드 집합을 확장한다. '논문군집기'는 top-down 방식과 bottom-up 방식을 이용하여 각 논문의 범주를 결정한다. 제안하는 시스템은 논문 검색 서비스와 같은 전문 분야에 대한 검색 서비스에 유용하게 사용될 수 있을 것이다.

키워드 : 결과 내 군집화, 단어군집화, K-means 알고리즘, 검색 엔진, 정보 검색

Abstract The search-results clustering problem is defined as the automatic and on-line grouping of similar documents in search results returned from a search engine. In this paper, we propose a new search-results clustering algorithm specialized for a paper search service. Our system consists of two algorithmic phases: Category Hierarchy Generation System (CHGS) and Paper Clustering System (PCS). In CHGS, we first build up the category hierarchy, called the Field Thesaurus, for each research field using an existing research category hierarchy (KOSEF's research category hierarchy) and the keyword expansion of the field thesaurus by a word clustering method using the K-means algorithm. Then, in PCS, the proposed algorithm determines the category of each paper using top-down and bottom-up methods. The proposed system can be used in the application areas for retrieval services in a specialized field such as a paper search service.

Key words : Search-Result Clustering, Word Clustering, K-means Algorithm, Search Engine, Information Retrieval

· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2009-0065895)

† 정 회 원 : 동아대학교 컴퓨터공학과
kmbae80@gmail.com
†† 학생회원 : 동아대학교 컴퓨터공학과
stfcap@gmail.com
††† 종신회원 : 동아대학교 컴퓨터공학과 교수
yjko@dau.ac.kr
jhkim@dau.ac.kr

논문접수 : 2009년 4월 17일
심사완료 : 2009년 12월 11일

정보과학회논문지: 소프트웨어 및 응용 제37권 제3호(2010.3)

1. 서 론

인터넷이 급격히 성장하면서, 사람들은 원하는 정보를 효과적으로 검색하기 위해 구글과 같은 검색엔진을 많이 이용한다. 하지만 검색 엔진은 단지 검색된 문서를 순위화해서 보여주기 때문에 사용자가 원하는 문서를 찾는 것은 쉽지 않다. 이를 해결하기 위한 하나의 방법은 결과 내 군집화 기법을 이용하는 것이다. 결과 내 군집화는 검색 엔진으로부터 검색된 결과 중에서 비슷한 문서를 자동으로 군집화해 줌으로써 사용자들이 원하는 문서를 효과적으로 찾을 수 있게 도와준다[1].

검색 엔진의 최근 경향 중 하나는 'Google scholar'와 같이 전문 분야에 대한 검색 서비스를 제공하는 것이다. 전문 분야에 대한 정보를 검색하는 대부분의 사용자는 자신이 원하는 분야가 고정되어 있고, 검색된 문서 중에서 관심 있는 분야와 주제가 일치하는 문서를 찾기 원한다. 그렇기 때문에 일반 문서 검색에서 보다 전문 분야에 대한 검색은 주제별로 결과를 군집화 할 수 있는 결과 내 군집화 기법이 더욱 필요하다.

검색 결과 내 군집화에 대한 많은 연구가 이루어지고 있으며, 대표적인 시스템으로 Scatter-Gather system[2], Grouper system[3], WebCat[4], RETRIVER[5], Lingo system[6,7] 등이 있다. 이 기법들은 검색된 결과 내에서 비슷한 문서들만 군집화하여 보여주기 때문에, 검색 결과 내에서 사용자가 원하는 분야에 일치하는 문서를 찾기 어렵다. Vivisimo system[8], FIHC system[9], SHOC system[10], HIGHLIGHT system[11], CIIRarchies system[12]은 이러한 문제를 해결하기 위해 검색 결과 내에서 주요 단어를 계층적으로 표현한 범주 체계를 이용하는 방법을 제안하였다. 하지만 생성된 범주 체계의 단어들은 일반적인 단어가 많기 때문에 전문 분야에 대한 검색 서비스에 적용하기 어렵다. 논문 검색 서비스와 같은 전문 분야에 대한 검색 서비스는 효과적인 검색 결과 내 군집화를 하기 위해서 전문 분야에 적합한 전문화된 범주 체계가 필요하다.

본 논문에서는 논문 검색 서비스에 전문화된 새로운 검색 결과 내 군집화 기법을 제안한다. 논문 검색 서비스에서는 검색 결과를 사용자가 원하는 연구 분야별로 잘 구분하는 것이 중요하다. 논문 검색 서비스를 이용하는 사용자는 대부분 자신이 연구하는 연구 분야가 있고, 그 분야에 대한 범주 체계를 잘 알고 있기 때문이다. 이를 위해, 본 논문에서는 논문이 가지는 좋은 특성인 논문의 제목과 고정된 연구 범주(category, topic)를 이용하여, 논문 검색 서비스에 전문화된 범주 체계를 생성하고, 생성된 범주 체계를 기반으로 각 논문들을 연구 범주별로 검색 전에 자동적으로 군집화하여 분류하는 알고리즘을 제안한다. 분류된 논문은 생성된 분류 정보를 사용하여 인덱싱된다. 사용자가 논문을 검색 하면 검색된 각 논문의 분류 정보를 이용해 검색 결과를 연구 범주 별로 분류하여 보여준다. 검색 결과는 연구 범주별로 분류되어 보여지기 때문에 사용자는 원하는 논문을 검색 결과 내에서 빠르고 쉽게 찾을 수 있다.

제안하는 시스템은 범주 체계를 생성하는 '범주체계생성기(Category Hierarchy Generation System)'와 논문을 군집화하는 '논문군집기(Paper Clustering System)'로 구성된다. 먼저, '범주체계생성기'에서는 검색된 논문을 분류할 수 있는 연구 범주 체계를 생성한다. 연구 범주 체계는 총 3개의 계층으로 이루어져 있다. 본 논문에서는 두 개의 계층(Lv1, Lv2)은 믿을 만한 기관에서 제공하는 연구 범주 체계를 기반으로 사람이 생성한다. 이를 위해, '한국과학재단(이하 KOSEF)'에서 제공하는 '과학기술연구분야분류' 연구 범주 체계를 이용한다. Lv1, Lv2 계층을 위한 키워드를 생성한 후, 단어 군집화 알고리즘을 이용해 자동으로 키워드를 확장한다. 그리고 '논문군집기'에서 제안하는 top-down방식과 bottom-up 방식을 이용하여 각 논문의 범주 체계를 결정한다. 본 논문의 구성은 2장에 검색 결과 내 군집화에 관련된 기존 연구들에 대해 살펴보고, 3장에 논문 검색을 위한 검색 결과 내 군집화 방법을 제안한다. 4장에서 제안한 System을 실험을 통해 평가하고 5장에서 결론을 낸다.

2. 관련 연구

이 장에서는 검색 결과 내 군집화에 대한 기존의 연구 방법들에 대해 알아본다. 검색 결과 내 군집화 기법은 크게 2가지 형태로 구분할 수 있다. 문서의 유사도를 기반으로 군집화하는 기법과 범주 체계를 이용해 문서를 범주 체계에 군집화하는 기법이 있다.

검색된 문서의 유사도를 기반으로 비슷한 문서들을 군집화하는 방법으로 다음과 같은 기법들이 연구되어 졌다. SCATTER/GATTER[2]는 검색 결과 내 군집화를 최초로 IR 시스템에 적용한 방법 중 하나이다. 하지만 온라인상의 웹(Web) 검색 엔진에는 적용하지 못하는 문제가 있다. Grouper[3]는 동적인 군집화를 이용해 웹(Web) 검색에 대한 결과를 군집화하여 보여주는 방법을 제안하였다. WebCat[4]과 RETRIVER[5]는 문서의 유사도만을 이용했을 때 생기는 단어 불일치 문제를 해결하기 위해 각각 K-means 군집화 알고리즘과 Fuzzy 군집화 알고리즘을 이용해 비슷한 문서들을 군집화하는 방법을 제안하였다. LINGO[6,7]는 SVD(Singular Value Decomposition)를 사용하여, 단어와 문서에서 의미가 있는 레이블(Label)을 추출하여 군집화하는 방법을 제안하였지만, 데이터의 양이 많은 경우 시간이

오래 걸리는 단점이 있다. 문서의 유사도를 기반으로 검색 결과 내 군집화를 하는 기존의 기법들은 비슷한 문서를 군집화해서 보여주기 때문에 서로 다른 주제를 가지는 문서들이 군집화되는 문제가 있다. 사용자는 원하는 주제에 대한 문서를 찾기 위해서는 문서의 순위화로 결과를 보여주는 검색 서비스와 같이 많은 문서를 확인해야한다. 특히 주제별로 군집화된 결과를 효과적으로 보여줘야 하는 전문 분야에 대한 검색 서비스에서는 좋은 결과를 보여주기 어렵다.

다음으로 범주 체계를 이용한 결과 내 군집화 기법들이 연구되어 졌다. Vivisimo[8]는 웹 검색 결과의 스니펫(snippet)에 존재하는 단어를 추출하여 범주 체계를 구성하는 방법을 제안하였다. FIHC[9]는 Frequent Itemsets Problem을 기반으로 범주 체계를 만들어 군집화하는 방법이다. SHOC[10]는 접미사 배열(Suffix Array)을 사용하여 문장을 추출하고, SVD(Singular Value Decomposition)를 통해 범주 체계를 구성한 방법을 제안했고, HIGHLIGHT[11]는 어휘 분석(Lexical Analysis)과 확률 모델을 사용하여 범주 체계를 구성하는 방법을 제안하였다. CIIRarchies[12]는 사전 계산된 언어모델(Pre-computed Language Model)을 사용하여 문장을 추출하고 재귀(Recursive) 알고리즘을 이용하여 범주 체계를 구성하는 방법을 제안하였다.

제안된 기법들은 문서에 존재하는 단어를 범주 체계 생성에 이용하였다. 하지만, 생성된 범주 체계는 문서에 존재하는 단어만을 이용하기 때문에 표준화된 범주 체계가 필요한 전문 분야 검색 서비스에는 적합하지 않다. 전문 분야에 대한 검색 서비스는 검색 결과를 범주 체계별로 군집화하기 때문에 믿을 만한 기관에서 만들어진 표준화된 범주 체계를 이용해야한다. 또한, 기존에 연구된 문서의 유사도를 기반으로 군집화하는 기법과 범주 체계를 이용해 문서를 범주 체계에 군집화하는 기법은 사용자가 검색 후 검색 결과에 대해 군집화를 진행하기 때문에 검색 후 결과를 보여주는 속도가 느리다는 단점이 있다.

본 논문에서는 논문 검색 서비스를 위한 결과 내 군집화를 위해 논문 검색 서비스에 적합한 전문화된 범주 체계를 자동으로 생성하고, 생성된 범주 체계를 이용해 각 논문을 범주 체계별로 분류하는 시스템을 제안한다. 제안된 시스템은 검색 전에 일괄처리방식으로 논문을 범주 체계에 군집화해서 각 논문의 분류 정보를 미리 결정한다. 검색 후에는 검색 결과들의 분류 정보만을 이용하여 분류함으로써 검색 후 군집화를 위해 필요한 계산 시간 및 계산량을 줄였다. 또한, 본 시스템은 표준화된 범주 체계별로 검색 결과를 군집화하여 보여 줌으로써 사용자가 원하는 논문을 검색 결과 내에서 쉽고 빠르게 찾을 수 있다는 장점이 있다.

3. 제안 시스템

제안하는 시스템은 범주 체계를 생성하는 '범주체계생성기(Category Hierarchy Generation System)'와 논문을 군집화하는 '논문군집기(Paper Clustering System)'로 구성된다. 그림 1은 시스템의 전체 구성을 나타낸다.

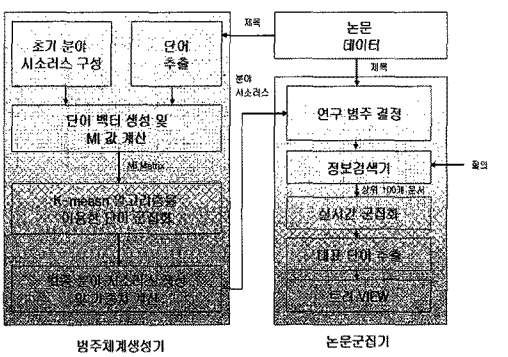


그림 1 시스템의 전체 구성

3.1 범주체계생성기(Category Hierarchy Generation System)

'범주체계생성기'는 '과학기술연구분야분류' 연구 범주 체계와 K-means 알고리즘 기반의 단어 군집화를 이용해 범주 체계에 필요한 단어를 생성한다. 범주 체계를 구성하는 계층적인 단어 집합을 분야 시소러스(Field Thesaurus)라고 부른다[12]. 분야 시소러스는 표 2와 같이 3개의 레벨들(범주(Lv1), 하위범주(Lv2), 키워드(Lv3))로 구성된다.

표 1 기존 연구와의 비교

	군집화 방법	군집화 계산 시기
SCATTER/GATTER[2], Grouper[3], WebCat[4], RETRIVER[5], LINGO[6][7]	문서의 유사도 기반	검색 후 실시간 계산
Vivisimo[8], FIHC[9], SHOC[10], HIGHLIGHT[11], CIIRarchies[12]	범주 체계와의 유사도 기반	검색 후 실시간 계산
제안 시스템	범주 체계와의 유사도 기반	검색 전 일괄 계산 검색 후 단순 분류

표 2 분야 시소러스 구성의 예

범주(category): Lv1	하위범주(subcategory): Lv2	키워드(keyword): Lv3
컴퓨터네트워크 (COMPUTER NETWORK)	인터넷 (INTERNET)	ISP
		TCP
		쇼핑몰 (SHOPPINGMALL)
	네트워크 (NETWORK)	패킷(PACKET)
		서브넷(SUBNET)

1. 범주(Lv1)와 하위범주(Lv2)는 KOSEF에서 제공하는 '과학기술연구분야분류' 연구 범주 체계를 이용해 생성한다.
2. 키워드(Lv3)는 하위범주(Lv2)에 있는 주제단어(subject word)와 관계된 단어로서 K-means 알고리즘을 이용해 생성한다.

3.1.1 범주와 하위 범주 생성

'범주체계생성기'에서 연구 범주 체계를 생성하기 위해 초기 분야 시소러스를 생성한다. 초기 분야 시소러스는 표 2의 '컴퓨터네트워크'와 같이 범주(Lv1)를 구성하는 단어와 하위범주(Lv2)를 구성하는 단어를 KOSEF에서 제공하는 '과학기술연구분야분류' 연구 범주 체계를 기준으로 사람이 직접 선정한다. 본 논문에서는 전기, 전자, 컴퓨터 영역에 적합한 범주와 하위범주를 초기 분야 시소러스로 구성하였다. 생성된 초기 분야 시소러스는 논문의 군집화에 필요한 범주와 하위범주를 대표하는 단어들로 구성된다. 각 범주와 하위범주를 대표하는 단어를 주제 단어(subject word)라 부른다. 본 논문에서는 각 논문의 하위범주를 적절히 군집화하기 위해 논문의 제목에 존재하는 단어를 이용한다. 논문의 제목에 존재하는 단어와 하위범주의 주제단어를 비교하여 일치하는 단어가 존재하는 하위범주로 군집화한다. 하지만 초기 분야 시소러스를 구성하는 주제단어는 하위범주를 대표하는 단어하나로만 이루어져 있기 때문에 정확한 군집화가 어렵다. 이를 해결하기 위해 본 논문에서는 하위범주를 대표하는 단어와 연관된 키워드 집합을 생성하여 논문을 군집화한다.

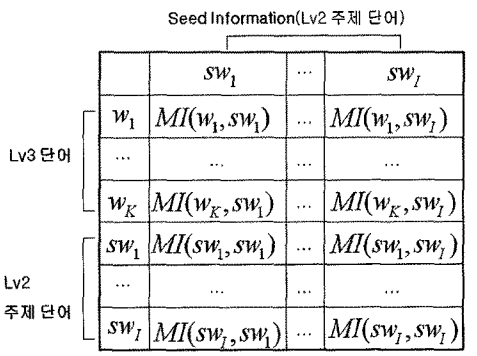
3.1.2 단어 군집화를 이용한 각 하위범주의 키워드 생성

분야 시소러스의 범주(Lv1)와 하위범주(Lv2)를 구성하고 난 후 분야 시소러스를 확장하기 위해 하위범주의 키워드 집합을 단어 군집화를 이용해 생성한다. 생성된 하위 범주의 키워드 집합은 분야 시소러스의 3번째 계층을 이루며 키워드(Lv3)라 부른다. 키워드는 범주와 하위범주와 다르게 '범주체계생성기'를 통해 자동으로 생성한다. 키워드 집합을 구성하기 위해 각 논문의 제목에 존재하는 모든 단어를 추출한다. 키워드 집합의 생성을 위한 후보 단어는 모든 분야에 일반적으로 쓰이는 단어이거나 각 분야별로 나오는 빈도가 너무 낮은 단어는 포함하지 않는다. 이를 위해, 단어가 존재하는 논문의 수가 3개 미만인 단어와 불용어를 제거한 단어만을 추출한다. 추출한 단어들을 K-means 알고리즘을 이용해 하위범주에 군집화시킨다. 하위 범주의 주제 단어(이하 Lv2 주제단어)를 K-means 알고리즘의 초기 중심벡터를 위한 씨앗정보(seed information)로 이용하여, 추출한 각 단어의 벡터와 Lv2 주제 단어의 벡터를 생성한다. 각 벡터는 Lv2 주제단어와 추출한 단어 사이의 상호정보(Mutual Information: MI) 값으로 표현된다. 본 논문에서 상호정보를 벡터의 값으로 이용하는 이유는 다음과 같다. 예를 들어, '인터넷'이라는 Lv2 주제단어가 있을 때 Lv2 주제단어의 하위단어는 인터넷 분야로 분류되는 논문에서 '인터넷'과 같이 자주 나오는 단어일 것이다. 두 단어가 같은 분야를 나타내는 단어라면 그 분야의 논문제목에서 같이 나올 확률이 높다고 볼 수 있다. 이 점을 고려해 본 논문에서는 각 논문 제목에서 두 단어가 같이 나타난 정보를 이용할 수 있는 상호정보(MI)를 이용하여 유사도를 계산하였다. 상호정보 값은 식 (1)을 이용해 계산한다.

$$MI(w_i, sw_j) = \log\frac{P(w_i, sw_j)}{P(w_i)\cdot P(sw_j)} = \log\frac{\frac{c(w_i, sw_j)}{N}}{\frac{c(w_i)}{N}\cdot\frac{c(sw_j)}{N}} = \log\frac{c(w_i, sw_j)}{c(w_i)\cdot c(sw_j)}N \quad (1)$$

$c(w_i)$는 i번째 추출된 단어의 빈도수, $c(sw_j)$는 j번째 주제 단어의 빈도수이고 $c(w_i, sw_j)$는 i번째 추출된 단어와 j번째 Lv2 주제단어가 같이 나타난 빈도수를 나타낸다. 각 빈도수는 전체 제목의 집합에서 발생한 빈도수를 계산한 것이다. 만약 추출된 단어의 전체 수가 K개이고 주제 단어의 전체 수가 I개 일 때 MI 값에 대한 행렬은 $|(K+I)*I|$의 크기를 가진다. 그림 2는 MI 행렬을 나타낸다.

그림 2에서 보는 것과 같이 모든 추출된 단어(Lv3 단어)는 Lv2 주제단어와의 MI값을 가지는 벡터로 표현된다. 그리고 Lv2 주제 단어의 각 벡터는 K-means 알고리즘의 초기 중심 벡터로 이용된다. 단어 군집화가 끝나면 추출된 단어는 각 하위범주에 군집화되어 진다. 최종적으로 하위범주에 군집화된 단어들은 하위 범주에 포함된 키워드로 선택된다. 본 논문에서 생성한 분야 시소러스는 새로운 논문이 추가되면 실시간 또는 주기적으로 새로 구축할 수 있다. 초기 분야 시소러스는 새로운 범주나 하위범주가 추가되지 않는 한 그대로 이용할 수 있고, 키워드로 선정할 단어 역시 이전에 추출한 단어를 저장해 놓으면 새로 추가된 논문의 제목에서만 단어를



	sw_1	...	sw_I
w_1	$MI(w_1, sw_1)$	...	$MI(w_1, sw_I)$
...	...	...	...
w_K	$MI(w_K, sw_1)$	...	$MI(w_K, sw_I)$
sw_1	$MI(sw_1, sw_1)$	...	$MI(sw_1, sw_I)$
...	...	...	...
sw_I	$MI(sw_I, sw_1)$	...	$MI(sw_I, sw_I)$

그림 2 상호정보(MI) 값 행렬

추출하면 되기 때문에 손쉽게 분야 시소러스를 새로 구축할 수 있다. 다만 새로 구축된 분야 시소러스를 기준으로 모든 논문을 다시 분류해야 하기 때문에 추가작업이 발생할 수 있다. 하지만, 제안하는 시스템은 사용자가 논문을 검색 했을 때, 실시간으로 논문의 분류 정보를 결정하는 것이 아니라 검색 전에 일괄처리방식으로 논문을 분류 해놓기 때문에 실시간 검색에는 장애가 되지 않는다. 이렇게 생성된 분야 시소러스는 '논문군집기'에서 논문의 분류 정보 생성에 이용된다. '논문군집기'는 하위범주에 군집화된 키워드와 논문의 제목에 존재하는 단어를 비교하여 각 논문을 적절한 하위범주로 분류한다. 생성된 키워드는 주제단어와 유사도 값을 가지고 있으며, 이는 논문 군집화 결과의 대표 단어를 선정하는데 이용된다.

3.2 논문군집기(Paper Clustering System)

'논문군집기'는 분야 시소러스와 제안하는 top-down, bottom-up 방법을 이용하여 각 논문의 범주 체계를 결정한다.

1. 각 논문은 top-down, bottom-up 방법을 이용해서 분야 시소러스의 범주들로 분류된다. 각 논문은 한 개 이상의 범주를 가질 수 있다.
2. 단계1에 의해 분류된 결과를 이용하여 검색 결과 논문을 하위 범주에 군집화한다.
3. K-means 알고리즘에 이용된 키워드의 중요도 값을 이용해 각 그룹(cluster)의 대표 단어를 선택한다.

3.2.1 논문의 범주 결정

논문의 범주 결정을 위해 범주 계층 생성기를 통해 생성된 분야 시소러스를 이용한다. '논문군집기'는 분야 시소러스의 하위범주의 주제단어와 키워드를 이용해 top-down, bottom-up 방법으로 각 논문의 범주를 결정한다. 논문의 범주를 결정하는 방법은 다음과 같다.

(1) top-down 방법

만약 논문의 제목에 하위범주의 주제단어가 직접 포함되어 있으면, 이 논문은 하위범주로 분류된다. 예를 들어, 논문의 제목에 '네트워크(NETWORK)'가 포함되어 있다면, 이 논문은 '네트워크' 하위범주로 바로 군집화된다[9].

(2) bottom-up 방법

만약 논문의 제목에 하위범주 주제단어가 포함되어있지 않지만 키워드단어(Lv3)가 포함되어 있다면, 그 논문은 키워드의 상위 범주인 하위범주(Lv2)로 분류된다. 예를 들어, 논문의 제목에 키워드인 '쇼핑몰(SHOPPING-MALL)'이 포함되어 있다면, 이 논문은 '쇼핑몰'의 상위범주인 '인터넷(INTERNET)'으로 분류된다.

사용자가 검색 엔진을 통해 질의를 입력하면, 검색 결과의 상위 100개의 문서를 순위화(ranking)해서 선정한다. 선정된 상위 100개의 문서는 각 논문의 군집화 정보를 기준으로 군집화하여 보여준다.

3.2.2 대표 키워드(Lv3) 선정

제안하는 시스템의 인터페이스는 그림 3에서 보는 바와 같이 계층 구조로 군집화된 결과를 보여준다. 범주(Lv1)과 하위범주(Lv2)의 주제단어는 결정되어 있지만, 마지막 레벨(Lv3) 주제 단어는 여러 개가 존재한다. 제안하는 시스템은 Lv3를 대표하는 주제단어를 선정하여 보여준다. Lv3를 대표하는 주제 단어는 하위범주 주제단어에 군집화된 논문의 제목에 존재하는 키워드들만 이용한다. 군집화된 논문의 제목에 존재하는 각 단어 중에서 하위범주 주제단어와의 유사도가 가장 높은 단어를 Lv3의 주제 단어로 선정한다. 유사도의 계산은 K-means 알고리즘을 이용한 단어 군집화에서 계산된 유사도를 그대로 이용한다(3.1.2절).

4. 실험 및 결과

4.1 실험 데이터

실험을 위해 전기, 전자, 컴퓨터 분야의 55,408개의

표 3 top-down , bottom-up 군집화 방법의 예

	제목	범주	적용방법
논문1	4G 네트워크에서 재구성성을 위한 개방형 API, (Open API for reconstruction in 4G NETWORK)	네트워크 (NETWORK)	Top-down
논문2	쇼핑몰 핵심 마케팅 전략 (SHOPPINGMALL marketing strategy)	인터넷 (INTERNET)	Bottom-up

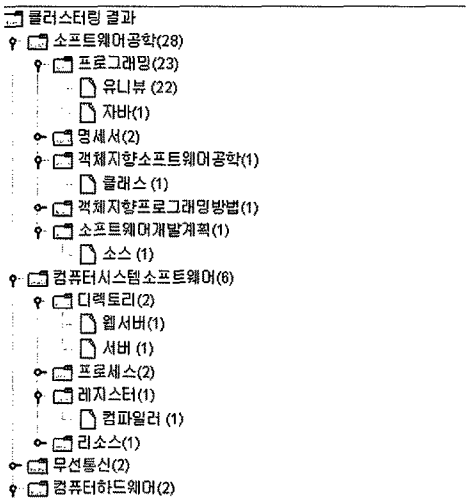


그림 3 '자바(Java)'를 질의로 입력했을 때의 결과화면

논문을 사용하였다. '범주체계생성기'를 통해 생성된 분야 시소러스는 21개의 범주(Lv1), 179개의 하위범주(Lv2), 8,463개의 키워드(Lv3)로 구성된다.

4.2 평가

일반적으로 검색 결과 내 군집화 기법은 다른 군집화 기법과의 비교가 어렵다. 그래서 제안하는 시스템을 평가하기 위해 생성된 키워드(Lv3)의 정확도와 논문 군집화의 정확도(accuracy)를 평가한다. 먼저 키워드(Lv3)의 정확도는 3명의 평가자가 각 하위범주의 키워드(Lv3)의 정확도를 평가한다. 키워드(Lv3)는 '범주체계생성기'를 통해 하위범주(Lv2)에 자동으로 군집화되었기 때문에 평가자는 하위범주에 군집화된 키워드(Lv3)가 적절하게 군집화되었는지 평가한다. 키워드(Lv3)의 정확도는 다음과 같이 계산한다.

$$\text{정확도}(Accuracy) = \frac{\text{정확히군집화된 } Lv3\text{키워드수}}{\text{전체 } Lv3\text{키워드수}} \quad (2)$$

다음으로 논문 군집화의 정확도는 다음과 같이 평가한다. 논문 군집화의 정확도를 계산하기 위해서 객관적인 기준이 필요하다. 본 논문에서는 논문 군집화의 정확도를 계산하기 위해 각 논문의 저자가 선정한 키워드(본 논문에서 생성한 분야 시소러스의 키워드(Lv3)가 아님)를 객관적인 기준으로 이용한다. 각 논문에 존재하는 키워드는 논문의 저자가 논문의 주제를 가장 잘 나타내는 단어를 선정하기 때문에 논문의 주제 또는 연구 분야를 분류하는 객관적인 기준이 될 수 있다. 이 점을 이용해 본 논문에서는 다음과 같이 논문 군집화 정확도를 평가하기 위한 정답셋을 생성한다. 정답셋은 저자의 키워드를 기반으로 신뢰할 수 있는 기관에서 제공하는 연구 범주 체계(분야 시소러스의 하위범주)로 분류하기 때문에 본 논문의 군집화 정확도를 평가할 수 있는 객관적인 기준이 된다[9].

1. 저자가 선정한 키워드가 존재하는 논문을 선정한다.
2. 선정된 각 논문의 키워드를 분야 시소러스에 있는 하위범주의 주제단어(Lv2 주제단어)와 비교한다. 이때 분야 시소러스의 범주, 하위범주는 KOSEF에서 제공하는 객관적인 연구 범주 체계이기 때문에 그대로 이용한다. 그리고 정답셋은 제안하는 시스템이 키워드(Lv3)의 존재 여부로 분류 정보를 구분하는 것과 달리, 키워드(Lv3) 생성을 위한 씨앗 정보로 이용된 하위범주의 주제단어만을 이용하여 분류한다.
3. 만약 키워드에 하위범주의 주제단어가 존재하면 논문의 연구 범주는 하위범주로 분류한다. 키워드에 하위범주의 주제단어가 존재하지 않는 논문은 정답셋에서 제외한다.
4. 연구 범주가 결정된 모든 논문을 정답셋으로 선정한다[9].

총 55,408개의 논문 중 저자의 키워드를 가지는 논문은 26,165개가 선정되었고, 분야 시소러스의 범주, 하위범주(Lv1, Lv2)만을 이용하여 연구 범주를 분류한 논문은 총 1,841개의 논문이 선정되었다. 26,165개의 논문 중 1,841개의 논문만이 분류 정보를 가진 이유는 저자의 키워드가 영어단어로만 존재하거나, 키워드로써의 역할을 못하는 일반 적인 단어를 키워드로 선정한 논문들이 많았기 때문이다. 그리고, 초기 분야 시소러스에 존재하는 키워드만 사용했기 때문에 1,841개의 논문만이 정답셋으로 선정되었다. 선정된 정답셋을 이용하여 본 시스템을 평가하였다. 논문의 군집화를 평가하기 위해 정확률(Precision), 재현율(Recall), F1-measure를 이용한다. 각 값은 아래 수식들과 같이 계산한다. 본 시스템의 평가는 검색 후 군집화에 대한 결과를 평가하는 것이 아니라, 검색 전 일괄처리방식으로 각 논문의 분류 정보를 생성하기 위해 진행되는 군집화에 대해 평가한다. 검색 후에는 검색된 논문들 중 같은 분류 정보를 가지는 논문별로 군집화하기 때문에 검색 후 군집화에 대한 평가는 하지 않는다.

$$\text{재현율}(\text{Recall}) = \frac{\text{제안한시스템기반으로분류된논문중정답셋에있고정확히분류된논문의수}}{\text{저자의키워드기반으로분류된논문(정답셋)의수}} \quad (3)$$

$$\text{정확률}(\text{Precision}) = \frac{\text{제안한시스템기반으로분류된논문중정답셋에있고정확히분류된논문의수}}{\text{제안한시스템기반으로분류된논문중정답셋에있는논문의수}} \quad (4)$$

$$F1-measuer = \frac{2*\text{Pr}*\text{Re}}{\text{Pr}+\text{Re}} \quad (5)$$

4.2.1 키워드(Lv3)의 평가 결과

키워드의 정확도는 2가지 형태로 측정하였다.

(1) 범주(Lv1)에 정확하게 군집화된 키워드의 비율

(2) 하위 범주(Lv2)에 정확히 군집화된 키워드의 비율

(1)은 키워드(Lv3)가 분야 시소러스의 가장 상위 레벨인 범주에만 정확히 군집화되었다고 판단했을 때의 비율이다. 다시 말해, 범주는 정확히 군집화되었지만, 하위범주는 틀리게 군집화된 경우도 정확히 군집화되었다고 판단했을 때의 비율을 나타낸다. (2)는 범주에 정확히 군집화되고, 하위범주에도 정확히 군집화되었다고 판단했을 때의 비율이다.

(1)과 (2)의 기준으로 본 시스템의 키워드의 군집화 정확도 결과는 그림 4와 같다. 그림 4에서 보는 바와 같이 (1)의 결과에서 단지 7.87%의 키워드가 범주와 하위범주 둘 다에 정확하게 군집화되지 않았다. 이 의미는 본 시스템에서 생성한 분야 분류 체계인 분야 시소러스의 키워드가 범주(Lv1)에 정확히 군집화된 비율이 92.13%라는 의미가 된다. 그리고, (2)의 결과에서 범주와 하위범주까지 모두 정확히 군집화한 비율은 85.09%이다. (1)의 결과는 본 논문에서 생성한 분야 시소러스를 이용해 21개의 클러스터(Lv1)에 8400여개의 단어들을 군집화했을 때의 결과이고, (2)의 결과는 (1)보다는 158개 많은 179개의 클러스터에 8400여개의 단어들을 군집화했을 때의 결과이다. 단어의 군집화 정확도에 대한 성능을 평가하기 어렵지만, 제안한 시스템을 이용해 8400여개의 단어를 179개의 많은 클러스터에 군집화했음에도 85.09%의 정확도를 보였기 때문에 본 논문에서 생성한 분야 범주 체계의 정확도가 높은 것을 확인할 수 있다. 키워드의 정확도는 분야 시소러스의 구축 정확도와도 연관이 있다. 키워드(Lv3)의 군집화 정확도는 초기 분야 시소러스(범주(Lv1), 하위범주(Lv2))를 어떻게 구성하느냐에 따라 영향을 받는다. 향후 각 분야의 전문가의 의견을 고려하여 초기 분야 시소러스를 구축한다면 더 높은 군집화 정확도를 얻을 수 있을 것이다. 키워드(Lv3)의 군집화 정확도가 향상된다면 부가적으로 논문 군집화의 정확도 역시 향상될 것이다.

4.2.2 논문 군집화의 평가 결과

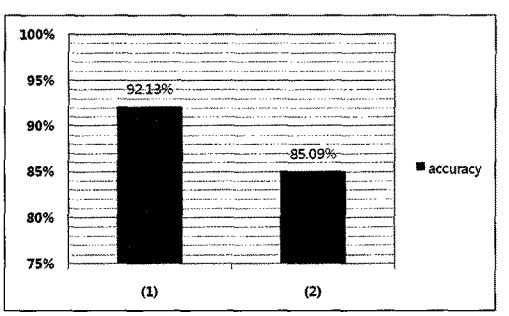


그림 4 키워드의 정확도

논문 군집화의 성능평가 역시 위와 같이 2가지 형태로 측정하였다.

(1) 범주(Lv1)에 정확하게 군집화된 논문의 비율

(2) 하위 범주(Lv2)에 정확히 군집화된 논문의 비율

(1), (2)의 의미는 키워드 정확도의 평가 방법과 같다. 그림 5에서 보는 바와 같이 범주와 하위 범주에 대해 각각 81.14%와 76.66%의 성능을 얻었다. 키워드(Lv3)의 군집화 정확도를 통해 분야 시소러스의 정확도가 대략 85% 정도라고 가정할 수 있다. 논문의 분류에 이용된 분야 시소러스의 정확도가 85%인 것을 감안한다면 제안한 논문의 군집화는 괜찮은 결과를 보였다고 판단할 수 있으며, 정확도가 높은 분야 시소러스를 이용한다면 더 나은 성능을 얻을 수 있을 것이다.

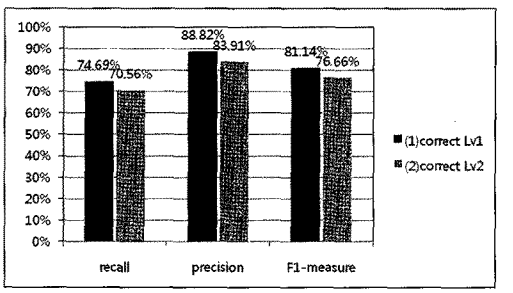


그림 5 논문 군집화의 성능

5. 결론 및 향후 과제

본 논문은 논문 검색 서비스를 위해 검색 결과를 효율적으로 군집화하는 새로운 알고리즘에 대해 제안하였다. 일반 문서 검색 서비스의 경우 검색대상이 되는 문서가 하나 이상의 주제를 포함하고 있기 때문에 검색 결과 내 군집화의 어려움이 있다. 하지만 본 논문에서는 논문이 주제의 변화가 적다는 특징과 논문이 가지는 메타정보(제목)를 이용하여 군집화에 필요한 군집화 정보를 사전에 생성하여 이용함으로써 효과적으로 검색 결과를 군집화할 수 있는 시스템을 제안하였다. 제안한 방법의 효율성은 키워드 정확도와 논문 군집화의 성능을 통해 입증되었다. 향후 과제로는, 정확률와 재현율을 향상시킬 수 있는 연구와, 논문 검색으로 제한된 검색 범위를 다양한 분야에 대한 검색에 적용시키는 연구가 필요할 것으로 생각된다.

참 고 문 헌

[1] G. Mecca, S. Raunich, A. Pappalardo, "A new algorithm for clustering search results," *Proc. Data & Knowledge Engineering*, pp.504-22, 2007.

[2] M. A. Hearst and J. O. Pedersen, "Reexamining

the cluster hypothesis: Scatter/gather on retrieval results," *Proc. SIGIR-96*, pp.76-84, 1996.

[3] O. Zamir and O. Etzioni, "Grouper: a dynamic clustering interface to Web search results," *Proc. Computer Networks: The International Journal of Computer and Telecommunications Networking*, pp.1361-1374, 1999.

[4] F. Giannotti, M. Nanni, and D. Pedreschi, "Webcat: Automatic categorization of web search results," *Proc. SEBD'2003*, pp.507-518, 2003.

[5] Z. Jiang, A. Joshi, R. Krishnapuram, and L. Yi, "Retriever: Improving web search engine results using clustering," *Proc. Managing Business with Electronic Commerce 02*, pp.59-81, 2002.

[6] S. Osinski and D. Weiss, "Conceptual Clustering using lingo algorithm: Evaluation on open directory project data," *Proc. IIPWM04*, pp.369-377, 2004.

[7] S. Osinski, J. Stefanowski, D. Weiss, "Lingo: Search results Clustering algorithm based on singular value decomposition," *Proc. the International Conference on Intelligent Information Systems (IIPWM)*, pp.359-368, 2004.

[8] P. Ferragina, A. Gulli, "A personalized search engine based on web snippet hierarchical clustering," *Proc. the World Wide Web Conference*, pp.189-225, 2005.

[9] B. Fung, K. Wang, and M. Ester, "Large hierarchical document clustering using frequent itemsets," *In SDM03*, 2003.

[10] D. Zhang and Y. Dong, "Semantic, hierarchical, online clustering of web search results," *Proc. The 3rd International Workshop on Web Information and Data*, pp.69-78, 2004.

[11] D. J. Lawrie and W. B. Croft, "Generating hiearchical summaries for web searches," *In SIGIR03*, 2003.

[12] Y. Wu and X. Chen, "Extracting features from web search returned hits for hierarchical classification," *Proc. International Conference on Information and Knowledge Engineering(IKE'03)*, pp. 103-108, 2003.

배 경 만

2004년 동아대학교 컴퓨터공학과 학사
2006년 동아대학교 컴퓨터공학과 석사
2008년 동아대학교 컴퓨터공학과 박사수료. 관심분야는 정보검색, 자연어처리, FAQ 시스템, 정보보호 등

황 재 원

2007년 동아대학교 컴퓨터공학과 학사
2009년 동아대학교 컴퓨터공학과 석사
관심분야는 자연어처리, 텍스트마이닝, 정보검색, 문서감정분류 등

고 영 중

정보과학회논문지 : 소프트웨어 및 응용
제 37 권 제 2 호 참조

김 종 훈

1974년 동아대학교 전자공학과 학사. 1975년~1977년 부산전자공업고등학교 교사 1977년 동아대학교 전자공학과 석사. 1986년 경북대학교 전자공학과 박사. 1986년~현재 동아대학교 컴퓨터공학과 정교수. 관심분야는 데이터마이닝, 정보보호 등