군집 중심 기반 문헌 검색 결과의 시각화

Visualization Method of Document Retrieval Result based on Centers of Clusters

지태창*, 이현진**, 이일병*
연세대학교 컴퓨터과학과*, 한국싸이버대학교 컴퓨터정보통신학부**

Tae-Chang Jee(garura@csai.yonsei.ac.kr)*, Hyunjin Lee(hjlee@mail.kcu.ac)**,
Yillbyung Lee(yblee@csai.yonsei.ac.kr)*

요약

기존의 문헌검색시스템은 검색 결과를 시각화하기 어렵기 때문에 문헌 제목과 검색어가 존재하는 부분에 대한 요약문을 보여주는 형태가 대부분이다. 이러한 방식은 문헌 검색 결과가 많은 경우 한 번에 문헌들을 살펴보는데 어려움이 있고, 문헌들간의 연관성을 알아보기 어렵다. 따라서, 본 논문에서는 웹 환경에 적합하도록 실시간으로 문헌 검색 결과를 시각화하는 방법을 제안하였다. 이를 위하여, 군집의 중심을 다차원 척도에 의해 저 차원 평면에 투사하는 단계와 오비탈 모형에 기반하여 개별 문헌들을 군집 중심을 기준으로 저 차원 평면에 표현하는 2단계 시각화 알고리즘을 제안하여, 문헌 군집의 관계를 쉽게 알아보고 개별 문헌들 사이의 유사성을 쉽게 확인할 수 있도록 하였다. 벤치마크 데이터와 실 데이터에 적용하여 실험하였으며, 실시간으로 검색 결과를 시각화 할 수 있다는 것을 실험을 통해 확인할 수 있었다.

■ 중심어 : | 시각화 | 문헌 군집화 | 실시간 문헌 검색 | 다차원 척도법 |

Abstract

Because it is difficult on existing document retrieval systems to visualize the search result, search results show document titles and short summaries of the parts that include the search keywords. If the result list is long, it is difficult to examine all the documents at once and to find a relation among them. This study uses clustering to classify similar documents into groups to make it easy to grasp the relations among the searched documents. Also, this study proposes a two-level visualization algorithm such that, first, the center of clusters is projected to low-dimensional space by using multi-dimensional scaling to help searchers grasp the relation among clusters at a glance, and second, individual documents are drawn in low-dimensional space based on the center of clusters using the orbital model as a basis to easily confirm similarities among individual documents. This study is tested on the benchmark data and the real data, and it shows that it is possible to visualize search results in real time.

■ keyword : | Visualization | Document Clustering | Real-Time Document Search | Multidimensional Scaling |

* 본 연구는 (주)윕스 및 산업자원부 연구과제로 수행되었습니다.

접수번호 : #070220-003
접수일자 : 2007년 02월 20일
심사완료일 : 2007년 03월 28일
교신저자 : 지태창, e-mail : garura@csai.yonsei.ac.kr

Ⅰ. 서론

21세기 지식 기반 사회에서는 빠른 속도로 증가하는 자료들을 어떻게 수집, 정리하고, 선별하여 유용한 정보만을 받아들이는가가 중요한 요소로 자리잡고 있다. 필요한 정보만 수집하는 방법 중의 하나로 인간의 인식 행위에 기반한 분류가 있다[1]. 지식 분류의 도구로서 군집화(Clustering)는 1960년대 후반에 연구가 시작되었으며, 1990년대에 들어 컴퓨팅 능력의 향상과 정보의 폭증에 힘입어 이에 대한 관심이 증대하였으며, 최근에는 이러한 군집화에 대한 지식 분류와 시각화에 대한 연구가 활발하게 진행되고 있다[2-5]. 군집화는 유사한 데이터 개체들의 집합인 군집 (Cluster)으로 데이터를 분할함으로써 데이터 속에 숨겨져 있는 의미 있는 정보를 자동으로 발견하는 것이다[6][7]. 이러한 군집화 알고리즘들의 장점 때문에 많은 문헌들에 숨겨져 있는 의미 있는 정보를 찾기 위하여 군집화 기법들을 적용하는 연구들이 SONIA, Scatter/Gather 등과 같은 프로젝트로, 또는 개별 연구도 활발하게 진행되고 있다[8-11].

현재의 문헌 분야의 군집화 및 정보검색 환경에서 사용자들은 일반적으로 텍스트 형태의 질의문을 입력하고, 텍스트 형태의 리스트로 검색결과를 제공받기 때문에 많은 검색결과로부터 사용자에게 적합하다고 판정되는 정보를 일일이 확인해야 하는 불편함이 있다. 이때 사용자의 적합성 판정은 검색결과의 표제나 저자, 목차, 요약문 등에 주로 의존하고 있다. 그러나 이러한 원문 대용물은 사용자가 원문의 내용을 이해하는 데 결정적인 역할을 하지 못한다는 연구 결과들이 보고되었다[12][13]. 따라서, 군집화 결과를 텍스트 형태로 제공하면 일반 사용자가 쉽게 군집화 결과를 일견할 수 없기 때문에 이를 보완하기 위한 시각화 방법에 대한 연구가 진행되고 있으나 미약한 실정이다[14].

온라인 정보 검색 시스템에서 정보를 검색할 때 사용자들은 빠른 시간에 결과를 얻기를 원하다. 그렇기 때문에 온라인 문헌 검색에 사용되는 문헌 군집화 시스템에서 군집화 시간이 오래 걸리면 실효성이 없어진다. 따라서, 군집화 결과를 시각화 하는데 있어서 빠른 시간에 처리해야 하는 것은 온라인 문헌 군집화 시스템의 필수 요소이다.

본 논문에서는 온라인 상에서 문헌 군집화 수행 후 군집화 결과를 시각적으로 표현하여, 일반 사용자가 좀 더 쉽게 군집화 결과와 문헌 간의 관계를 알 수 있는 알고리즘을 제안하려고 한다. 제안하는 방법은 기존의 다차원 데이터를 저 차원으로 축소하는 방법이 아닌 군집화 결과를 직접 이용하는 알고리즘이기 때문에 군집화 결과를 그대로 유지하면서 시각화 수행 시간에서 빠른 결과를 기대할 수 있다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 기존의 군집화 결과 시각화 방법들에 대해서 기술하고 3장에서는 제안하는 시각화 방법을 설명한다. 그리고 4장에서는 제안한 방법을 이용한 실험결과를 분석함으로써 그 유효성을 보인다. 마지막으로 5장의 결론에서는 결과 및 향후 연구 과제를 제시한다.

Ⅱ. 관련 연구

군집화의 결과를 인간이 인식하기 쉬운 2차원 평면으로 시각화 하기 위해서 고차원의 데이터를 저차원으로 축소하는 방법이 필요하며 이를 위해 PCA(Principal Components Analysis), 다차원 척도법 (Multidimensional Scaling: MDS), SOM(Self Organizing Map)과 같은 위상 보존 알고리즘 (Topology preserving algorithms)들이 사용된다[15]. 위상 보존 알고리즘은 고차원 데이터공간의 데이터 구조를 최대한 보존하여, 저 차원의 공간에서 표현하는 방법이다. 이 방식은 "근거리 점들을 근거리 점들로 (또 때에 따라선 원거리 점들을 원거리 점들로) 표시하는 것과 같이 한 공간의 점들을 다른 공간의 점들"[16] 로 표현한다.

Galaxies[14][17]는 문헌들의 고차원 발현을 2차원 산포도(scatterplot)로 낮춤으로써 군집과 문헌 밀접성을 드러내도록 시각화 하였다. 이 방법은 문헌들이 유클리디언(Euclidean) 거리나 코사인(Cosine) 값과 같은 유사성 척도(similarity metric)를 통해 고차원의 공간에서 응집된 후, 군집 중심으로 문헌 군집을 반영하는 2D

공간으로 투사된다. ThemeScape[17]에서는 서로 다른 두 가지의 차원감소 기술이 적용되었다. 대형 문헌군들에는 스트레스 최소 고정형 (Anchored Least Stress) 알고리즘이 사용되고, 1,500개 이하의 소형 문헌군들에는 셰파드 다차원 척도(Shepard MDS) 알고리즘이 사용되었다. 이 연구에서 기준 평면은 문헌군을 투사하는데 사용되었고, 미가공 문헌군에서 발견할 수 있는 것처럼 봉우리는 대량의 문헌 군집, 계곡은 문헌 군집 사이의 거리를 나타내도록 시각화 하였다.

문헌 집합의 쌍방향 연구를 위하여 WEBSOM 프로젝트[18]와 여러 연구들에서는 SOM을 활용하였다[19-21]. SOM은 문헌 집합의 통찰력있는 시각을 제공하는 맵상에 문헌을 재현하는데 이용된다. 온전한 WEBSOM 방식은 어휘분류맵과 문헌맵으로 구성된 2단계 SOM 구조가 필요하나, 이 연구에서 SOM은 어휘분류맵을 구축하는데 이용되었다. 먼저 밀접한 관계의 어휘들이 맵상에서 서로 근접하게 나타나고, 어휘분류맵상에 그 텍스트를 위치시켜 부호화하여, SOM 알고리즘으로 어휘분류맵 공간에서 문헌벡터를 사용하여 문헌맵을 생성하였다.

III. 제안하는 방법

문헌 군집화를 위해서는 문헌을 군집화 알고리즘이 이해할 수 있는 형태로 변환해야 하는데, 본 논문에서는 문헌-특징 벡터를 사용하였다. 문헌-특징 벡터는 문헌에 대해 형태소 분석과 파싱을 거쳐 단어들을 추출해 내고, 이 단어가 각 문헌에 나타나는 빈도수를 조사해서 이 단어별 빈도수를 문헌의 특징 벡터로 구성하는 방식이다[22]. 문헌-특징 벡터는 특징 벡터의 차원이 수백에서 수천 차원을 이루고 있어서 실시간 시각화에 있어서 문제가 된다.

본 논문에서는 데이터 량이 증가하거나 문헌-특징 벡터의 차원이 증가하더라도 시각화할 수 있는 선형의 계산시간을 갖는 알고리즘을 제안한다.

1. 문헌 군집화 구성도

제안하는 문헌 군집화의 구성도는 [그림 1]과 같다. 우선, 전체 문헌 또는 검색된 문헌에 대해 형태소 분석 및 파싱을 적용해서 문헌-특징 벡터를 구성한다. 문헌-특징 벡터는 군집화 알고리즘의 입력이 되어 군집화가 이루어 진다. 본 논문에서는 군집화 알고리즘으로 K-Means[7]를 적용하였다.

K-Means에 의해 군집화가 이루어지고 나면 생기는 정보는 각 군집의 중심점과 각 문헌들의 개별 군집에 대한 소속도이다. 이 정보들을 이용하여 2단계 시각화가 이루어지게 된다. 우선, 1단계로 각 군집의 중심점들의 거리 정보에 다차원 척도법을 사용해서, 다차원 군집 중심점을 저 차원 평면상에 사상한다. 2단계에서는 군집 중심들과 개별 문헌의 각 군집에 대한 소속도를 이용하여 개별 문헌들을 저 차원 평면상에 표현한다.

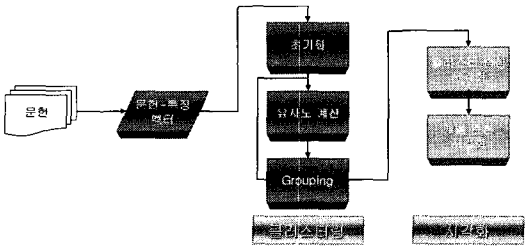


그림 1. 문헌 군집화 구성도

2. 클러스터 중심 시각화

본 논문에서는 다차원 척도법에 의해 군집의 중심을 저 차원 평면에 사상한다. 군집화가 완료되면, 처음에 결정한 숫자만큼의 군집 중심이 생기고, 군집 중심간의 거리를 계산할 수 있다. 이 군집 중심간의 거리를 다차원 변환을 통하여 저 차원 평면에 사상한다. [표 1]은 12개의 군집 중심간의 거리를 계산한 것이고, [그림 2]는 [표 1]의 값을 다차원 척도법을 이용하여 2차원 평면상에 표상한 것이다.

군집의 거리는 코사인 상관도로 계산해서 1이면 가깝고, 0이면 더 먼 값 이지만, 다차원 척도법은 0이 가깝고, 1이 먼 값을 가져야 하기 때문에 군집의 거리에 코사인의 역함수(arc cosine)를 적용하여 [표 1]을 구하였

다. [그림 2]를 보면, 가장 가까이 있는 군집은 0-1과 3-11 군집이고, 이 두 쌍의 역 코사인 상관도는 [표 1]에서 각각 1.09, 1.27로 다른 쌍에 비해서는 작은 값으로, 고차원 공간에서도 가까이 있다는 것을 확인할 수 있다. 하지만, 4-7은 역 코사인 상관도가 1.06으로 위 두 군집보다 작은 값으로 고차원 공간에서는 더 가까이 있는 것으로 계산되었지만, [그림 2]에서는 훨씬 멀리 존재하는데 이는 다른 군집들과의 관계를 같이 고려하여 표현했기 때문이다.

표 1. 12개 군집 중심의 거리

	0	1	2	3	4	5	6	7	8	9	10	11
0	-	1.09	1.44	1.13	1.36	1.05	1.15	1.12	1.19	1.41	1.37	1.20
1	1.09	-	1.51	1.28	1.41	1.18	1.27	1.28	1.36	1.49	1.48	1.36
2	1.44	1.51	-	1.46	1.51	1.46	1.53	1.40	1.50	1.39	1.44	1.50
3	1.13	1.28	1.46	-	1.47	1.33	1.31	1.32	1.39	1.45	1.42	1.27
4	1.36	1.41	1.51	1.47	-	1.38	1.45	1.06	1.43	1.50	1.45	1.46
5	1.05	1.18	1.46	1.33	1.38	-	1.25	1.25	1.32	1.43	1.39	1.32
6	1.15	1.27	1.53	1.31	1.45	1.25	-	1.31	1.34	1.48	1.48	1.20
7	1.12	1.28	1.40	1.32	1.06	1.25	1.31	-	1.27	1.36	1.34	1.27
8	1.19	1.36	1.50	1.39	1.43	1.32	1.34	1.27	-	1.40	1.41	1.28
9	1.41	1.49	1.39	1.45	1.50	1.43	1.48	1.36	1.40	-	1.38	1.37
10	1.37	1.48	1.44	1.42	1.45	1.39	1.48	1.34	1.41	1.38	-	1.40
11	1.20	1.36	1.50	1.27	1.46	1.32	1.20	1.27	1.28	1.37	1.40	-

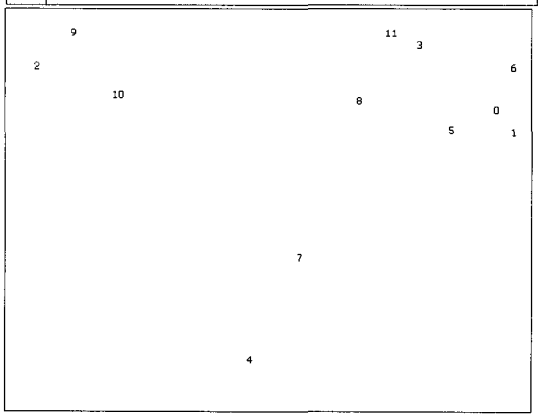


그림 2. 12개의 클러스터 중심에 다차원 척도법을 적용한 결과

3. 개별 문헌 시각화

본 논문에서는 개별 문헌 시각화에 원자 구조에 관한 물리학적 원리인 오비탈(Orbital)의 정리를 모델링한 알고리즘을 제안하였다.

원자는 전자와 원자핵으로 이루어져있고, 그 밖에 양성자와 중성자도 존재한다[23]. 전자는 (-) 전하를 띤 입자이며, 이 입자의 질량당 전하량은 항상 1.76×108C/g이다. 원자는 대부분 빈 공간이고, 중심부에 양전하를 띤 원자 질량의 대부분을 차지하는 곳이 있고, 이것이 원자핵이다. 원자 모형은 실험을 통해서 점점 개념을 발전해 왔는데, 1807년의 돌턴은 원자 모형이 딱딱한 공모양이라고 했고, 1903년 톰슨은 건포도가 든 푸딩 모형으로 정의하였다. 1911년 러더퍼드는 행성 모형으로 정의하였고, 1913년 보어는 궤도 모형을 정의하였다. 간단하게 원자 모형을 설명할 때는 보어의 모형을 많이 사용한다. 1926년에는 [그림 3]과 같은 오비탈(orbital) 모형이 정의되었는데, 이는 양자 역학에 토대를 두어 만들었고, 전자는 원자핵 주위에 구름처럼 퍼져 있으며, 어느 공간에서 전자를 발견할 수 있는 확률을 계산하여 확률 분포를 구름처럼 그린 모형으로 전자 구름 모형이라고도 불린다.

원자 모형에서 중요한 것은 원자핵과 전자의 위치이다. [그림 3]의 오비탈 원자 모형에서는 원자핵은 3좌표축이 교차하는 지점이 되고, 전자는 그 주위의 구름처럼 퍼져있는 곳 중의 어느 하나가 된다. 제안하는 방법에서는 군집화 결과를 오비탈 모형에 대응하여, 군집 중심은 원자핵과 대응되고, 개별 문헌은 전자에 대응된다. 이 모형을 그림으로 나타내면, 군집 중심은 다차원 척도법에 의해 계산된 저 차원 평면상의 좌표가 되고, 그 주위에 개별 문헌이 구름 형태로 분포하게 된다.

그림 3. 오비탈 원자 모형

하지만, 이처럼 개별 문헌이 확률에 의해서 모호하게 존재한다면, 문헌 시각화의 주 목적인 문헌간의 관계를 파악하는 것은 어렵다. 따라서, 개별 문헌도 존재할 확률이 아닌, 관찰 시점의 위치로 표현할 수 있어야 사용자에게 명확한 관계를 제시할 수 있다. 개별 문헌의 위치는 다른 군집 중심과의 관계로 계산한다. 군집 중심을 원자핵이라고 했으므로 (+) 전하를 가지게 되고, 개별 문헌을 전자라고 했으므로 (-) 전하를 가지게 된다.

따라서, 전자인 개별 문헌들은 자신이 속한 군집의 원자핵(군집 중심) 뿐만 아니라, 자신이 속하지 않은 군집의 원자핵(군집 중심)에도 영향을 받게 된다. 군집의 원자핵들은 다차원 척도법에 의해 특정 위치에 지정되기 때문에 그에 따른 전자들의 위치 계산도 가능하게 된다.

제안하는 군집화 결과 시각화 알고리즘은 다음과 같다.

우선, 3.2절과 같이 다차원 척도법을 이용하여 군집 중심의 좌표 CC_i를 계산한다. 여기서, i는 $i \in \{1...K\}$이고, K는 군집의 개수이다.

다음, 군집에 속한 개별 문헌의 좌표를 문헌벡터로 계산한다. 문헌벡터는 크기 $\| \overrightarrow{C_i d_k} \|$ 와 방향 $\overrightarrow{C_i d_k}$으로 이루어져 있다. 여기서, $\overrightarrow{C_i d_k}$는 i번째 군집 C_i의 중심 좌표 CC_i에서 k번째 문헌 d_k까지의 벡터이고, k는 $k \in \{1...N_i\}$이고, N_i는 i번째 군집에 속한 문헌의 개수이다.

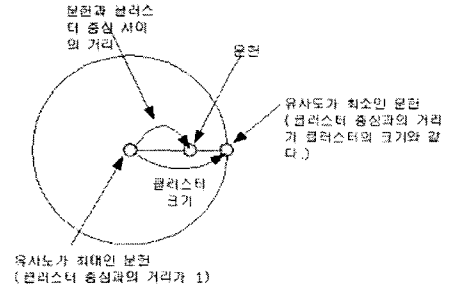


그림 4. 클러스터내의 문헌의 위치 예

문헌벡터의 크기는 문헌과 소속 군집 중심의 거리를 이용하여 계산한다.

$$\| \overrightarrow{C_i d_k} \| = SC_i \times coef_{c_i}(d_k) \quad (1)$$

여기서, SC_i는 i번째 군집의 크기이고, $coef_{c_i}(d_k)$는 k번째 문헌 d_k의 i번째 군집 C_i에 대한 소속도고, 이 소속도는 코사인 상관계수로 구한다. 코사인 상관계수는 0이 소속도가 작고, 1이 소속도가 크기 때문에 [그림 4]와 같이 문헌벡터는 소속도가 클수록 군집 중심과 가까이 존재하게 된다. 군집의 크기는 다음과 같이 계산한다.

$$SC_i = \frac{\min_{j=1, j \neq i}^{K} (CC_i - CC_j)}{2} \quad (2)$$

본 연구는 시각화를 목적으로 하기 때문에 군집들끼리 서로 겹치면, 문헌이 어느 군집에 속해 있는지 구별하기 어려워지고, 군집들끼리 너무 떨어지게 하면, 군집의 크기가 작아져서 문헌들 간의 구별이 어렵게 된다. 따라서, 군집과 문헌들을 서로 잘 분리할 수 있는 크기로 최근접 군집 사이의 거리로 군집의 크기를 결정하였다.

문헌벡터의 방향은 문헌과 소속하지 않은 군집들과의 관계로 계산한다. 즉, i번째 군집 C_i에 소속된 문헌 d_{k_1}과 d_{k_2}가 있을 때, 문헌 d_{k_1}과 d_{k_2}가 군집 C_i에 대한 소속도가 동일하더라도, 문헌 d_{k_1}이 군집 C_{j_1}과 소속도가 크고 군집 C_{j_2}와 소속도가 작은 반면, 문헌 d_{k_2}가 군집 C_{j_1}과 소속도가 작고 군집 C_{j_2}와 소속도가 크다고 하자. 이 경우 문헌 d_{k_1}은 군집 C_{j_1}과 가까운 방향으로 향하게 되고, 문헌 d_{k_2}는 군집 C_{j_2}와 가까운 방향으로 향하게 된다. 이와 같이 문헌벡터는 개별문헌이 소속된 군집뿐 만 아니라 소속하지 않은 군집들과의 소속도에 의해 방향이 결정된다. 문헌벡터의 방향을 결정하기 위해서는 먼저 군집 중심 사이의 벡터방향을 결정한다.

$$\overrightarrow{C_i C_j} = \frac{(x_j - x_i, y_j - y_i)}{\| \overrightarrow{C_i C_j} \|} \quad (3)$$

$\overrightarrow{C_i C_j}$는 i번째 군집에서 j번째 군집으로의 벡터방향이고, 여기서, j는 $j \in \{1..k\}, j \neq i$이고, x_i는 i번째 군집 중심 CC_i의 x축 좌표이고, y_i는 i번째 군집 중심 CC_i의 y축 좌표이다. 현재는 2차원 평면을 고려하기 때문에 x, y두 좌표인 것이고, 3차원 인 경우에는

x, y, z 좌표로 계산하면 된다. $\| \overrightarrow{C_iC_j} \|$ 는 i번째 군집 중심 CC_i와 j번째 군집 중심 CC_j 사이의 거리로, $\overrightarrow{C_iC_j}$를 거리가 1인 단위벡터로 만들어 준다.

문헌벡터의 방향은 다음과 같이 계산한다.

$$\overrightarrow{C_id_k} = \sum_{j=1, j\neq i}^{K} \left(coef_{C_i}(d_k) \times \overrightarrow{C_iC_j} \right) \quad (4)$$

[그림 5]는 제안하는 알고리즘을 그림으로 표현한 것이다. (a)는 수식 (1)을 도식한 것으로 문헌 d_1과 d_2는 군집 C_1에 속해있고, 소속도가 같아서 같은 거리에 위치해 있다. 점선은 문헌이 존재할 가능성이 있다는 것을 의미하고, 현재는 점선 위에 동일한 확률로 존재한다. (b)는 수식 (3)을 도식한 것으로 군집 C_1과 다른 군집들과의 단위방향벡터들을 생성한다. (c)는 수식 (4)를 도식한 것으로 단위방향벡터와 소속도를 이용하여 문헌의 위치를 결정한 것이다. 문헌 d_1은 군집 C_3와 가깝고 군집 C_5와 먼 경향을 보이고, 문헌 d_2는 군집 C_2, C_5와 가깝고, C_3, C_4와 먼 경향을 보이는 것을 알 수 있다.

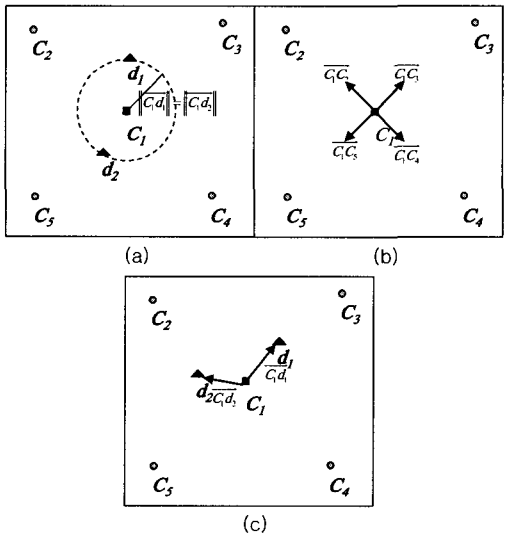


그림 5. 제안하는 알고리즘의 도식도

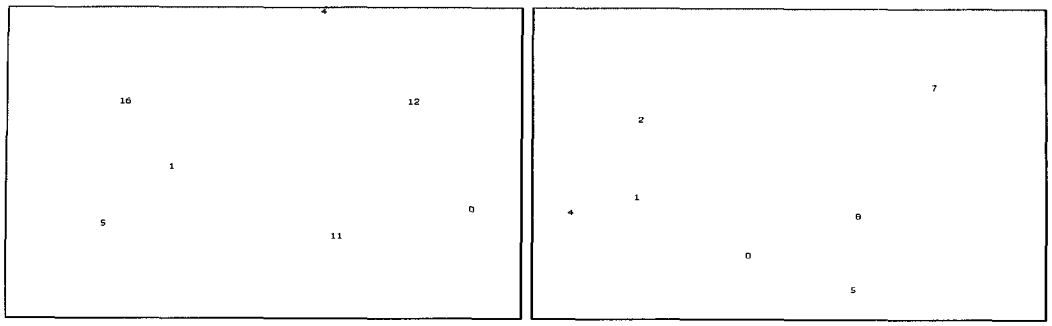
IV. 실험 결과

실험은 Intel CPU 2GHz, RAM 1GB의 시스템에서 C# 언어로 구현하였다. 사용된 컴파일러 버전은 .Net Framework 1.1이다. 실험 데이터는 실제 문헌 데이터를 사용하였는데, Reuter-21578 문헌 집합(Reuter)[24]에서 선택한 4개의 데이터 집합과 특허 문헌[25]에 대하여 각각 Car Navigation, Computer, Iris Recognition, Printer를 검색한 결과(Patent)를 이용하였고, 이 문헌들에 대한 통계 정보는 [표 2]와 같다.

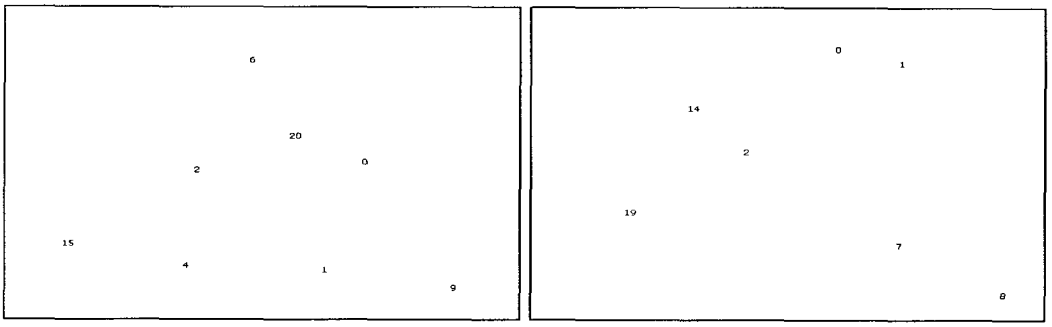
표 2. 실험에 사용된 데이터

Data Sets		*Num of instances*	*Num of features*	*Num of clusters*
Reuter	February	132	250	7
	April	2,106	2,671	7
	June	988	1,740	8
	October	605	828	7
Patent	Car Navigation	1,000	765	6
	Computer	1,000	739	5
	Iris Recognition	1,000	736	7
	Printer	1,000	642	6

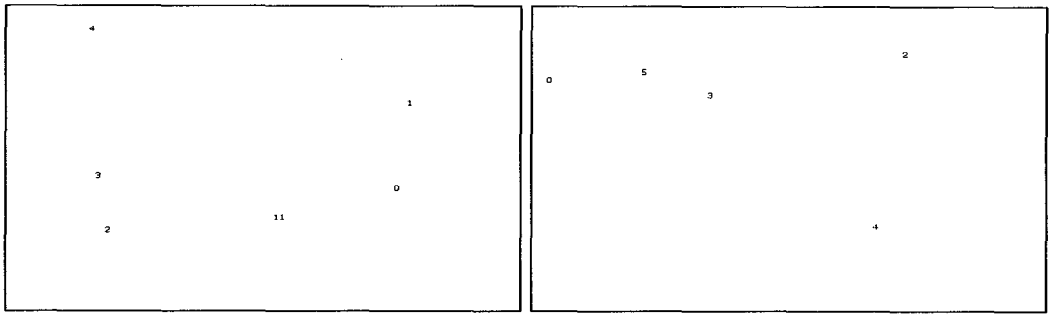
[그림 6]은 다차원 척도법에 의해서 군집 중심들을 2차원 평면상에 나타난 결과이다. (a), (b), (c), (d) 는 Reuter 데이터에 대한 결과이고, (e), (f), (g), (h) 는 Patent 데이터에 대한 결과이다. 수행 시간은 Reuter의 경우 (a) February는 0.015초, (b) April은 0.082초, (c) June은 0.042초, (d) October는 0.032초가 걸렸고, Patent의 경우 (e) Car Navigation, (f) Computer, (g) Iris Recognition, (h) Printer 모두 0.042초가 걸렸고, Reuter의 경우 군집 수가 많은 (b) April의 경우에 수행 시간이 좀 더 걸리고, 나머지는 거의 비슷한 시간이 소요되는 것을 확인 할 수 있다.



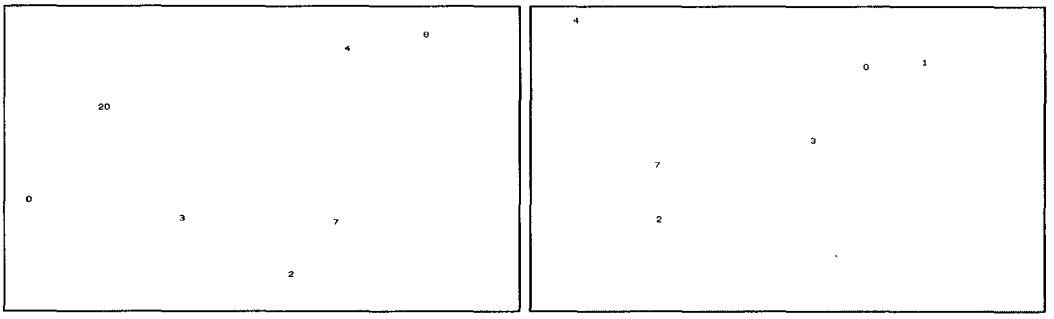
(a) Reuter - February (b) Reuter - April



(c) Reuter - June (d) Reuter - October

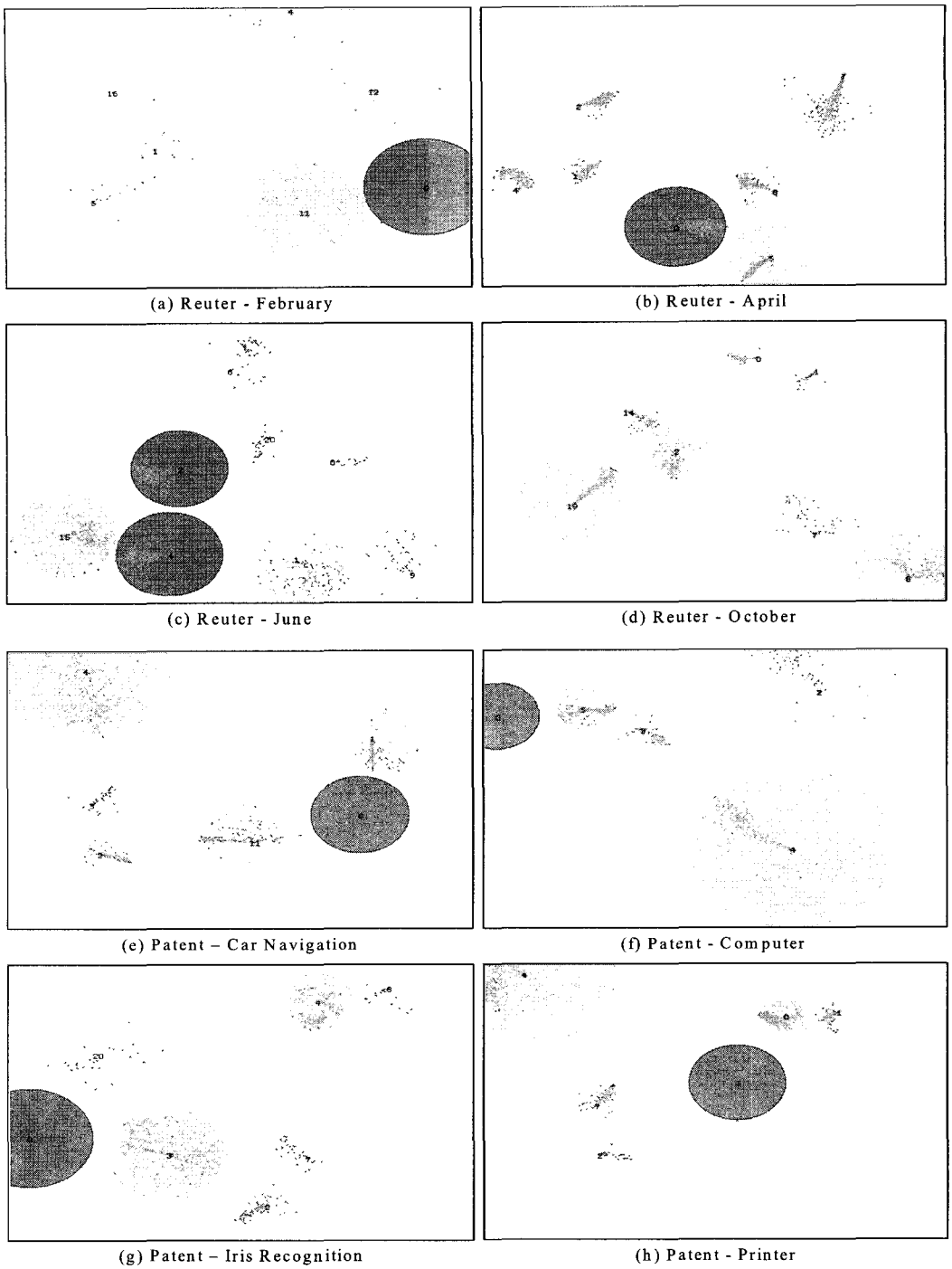


(e) Patent – Car Navigation (f) Patent - Computer



(g) Patent – Iris Recognition (h) Patent - Printer

그림 6. MDS를 이용하여 클러스터 중심을 2차원 평면에 배치한 결과



(a) Reuter - February (b) Reuter - April

(c) Reuter - June (d) Reuter - October

(e) Patent – Car Navigation (f) Patent - Computer

(g) Patent – Iris Recognition (h) Patent - Printer

그림 7. 제안하는 방법으로 개별 문헌들에 대한 시각화 처리 결과

[그림 7]은 제안하는 방법으로 개별 문헌들을 시각화 한 결과이다. (a), (b), (c), (d)는 Reuter 데이터에 대한 결과이고, (e), (f), (g), (h)는 Patent 데이터에 대한 결과이다. 수행 시간은 (b) Reuter - April의 경우 0.002초가 소요되고, 다른 데이터 집합에 대해서는 0.001초가 소요되었다. 수행 시간을 보면 제안하는 방법은 개별 문헌을 처리해야 하기 때문에 데이터 량이 100건 ~ 2,000건이고, 다차원 척도법은 문헌 중심을 처리해야 하기 때문에 데이터 량이 10건 이내로, 제안하는 방법이 데이터 량은 더 많지만 수행시간은 더 적게 소요되었다.

[그림 7]을 보면, 개별 문헌들의 분포가 어느 한쪽 방향으로 쏠리지 않기 때문에 개별 군집에 속한 문헌들이 서로 완전히 같은 경향을 보이지는 않는 다는 것을 확인할 수 있고, 같은 군집에 속한 문헌들일지라도 더 비슷한 문헌과 비슷하지 않은 문헌들이 섞여 있다는 것을 알 수 있다. 따라서, 사용자가 특정 문헌을 선택했을 때 그 문헌과 유사한 문헌들을 좀 더 정확하게 알 수 있게 된다는 장점이 있다.

[그림 7]의 군집 내의 색이 다른 것은 군집에 속한 문헌들의 개수를 비교하기 위함이다. 즉, 색이 진할 수록 더 많은 문헌들이 군집에 소속되어 있고, 색이 옅을 수록 더 적은 문헌들로 구성되어 있다. 전체 문헌에 대해 10% 단위로 표현하였다.

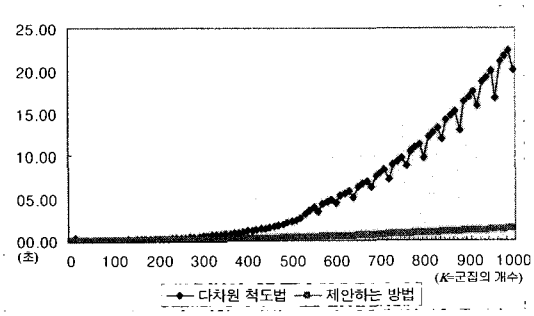


그림 8. 다차원 척도법과 제안하는 방법의 수행 시간 비교

다차원 척도법의 수행시간의 시간복잡도는 $O(lK^2)$이다. 여기서 K는 군집의 개수이고, l은 다차원 척도법 알고리즘의 반복 횟수이다. 즉, 군집의 개수가 증가할수록 그 제곱에 비례하여 시간이 걸리고, 안정화되기 위한 반복 수행 시간도 걸리게 된다. 제안하는 방법의 시간복잡도는 $O(KN)$으로, 여기서 K는 군집의 개수이고 N은 문헌의 개수이다. 즉, 문헌이 증가하거나, 군집이 증가하면, 시간은 그에 따라 선형적으로 증가한다. 실제 수행 시간에 대한 실험 결과는 [그림 8]과 같다. 여기서, 문헌의 수는 1,000으로 고정하고, 군집의 수를 10 ~ 1,000 까지 증가하면서 수행 시간을 측정하였다. 다차원 척도법은 군집 증가량의 제곱에 비례하여 시간이 증가하고, 제안하는 방법은 군집의 증가량에 선형적으로 증가하는 것을 알 수 있다. 결국, 1,000개의 문헌을 시각화할 때 다차원 척도법 만 사용하여 문헌을 표시하기 위해서는 23초가 걸리지만, 다차원 척도법과 제안하는 방법을 결합하여 다차원 척도법으로 군집 중심 $K=10$을 표현하고, 제안하는 방법으로 문헌 1,000개를 표현하면 1초 이내가 걸리게 된다. 즉, 제안하는 방법이 온라인 시스템에 더 적합하다고 할 수 있다.

V. 결론 및 향후 연구 과제

본 논문에서는 문헌을 검색 한 결과에 대해 군집화가 이루어져 문헌을 군집화 한 후, 이 결과에 대해 다차원 척도법을 이용하여 군집 중심을 저 차원 평면에 배치하고, 개별 문헌들을 군집 중심을 기준으로 표현하는 시각화 방법을 제안하였다.

시각화를 완성하기 위한 첫 번째 과정으로 개별문헌에 나타한 용어를 가지고 군집화 과정을 수행하였다. 두번째 과정에서는 개별문헌 시각화 방법을 설계하였는데, 첫 번째 과정에서 생성된 군집 중심 데이터에 다차원 척도법을 적용하여 저 차원 공간에 군집 중심을 배치하였고, 군집 중심과 개별 문헌들과의 관계를 오비탈 원자 모형으로 설명하여 개별 문헌들을 2차원 공간에 표현하게 설계하였다. 이로써 사용자가 유사한 문헌들인 군집들의 관계를 파악할 수 있을 뿐만 아니라 개별 문헌들 사이에서도 유사성을 쉽게 확인할 수 있다. 또한 실험을 통하여 온라인 검색 시스템에 적용할 수 있는 것을 확인하였다.

향후 연구과제는 다음과 같다. 본 논문에서는 군집과

개별 문헌들을 2차원 평면에 표현하였지만, 다차원 척도법에 의해 축소되는 차원에는 제한이 없기 때문에 군집 중심을 3차원에 표시하면, 개별 문헌들도 쉽게 3차원으로 표현할 수 있다. 따라서, 3차원 시각화 방법이 필요하며 이런 3차원 방법이 온라인 검색 시스템에 적합한지는 실험을 통해 검증해 보아야 한다.

참 고 문 헌

[1] A. Toffler, *The Third Wave*, New York, Bantam Books, 1990.

[2] K. Cox, et al. "A Multi-Modal Natural Language Interface to an Information Visualization Environment," IJST, Vol.4, pp.297-314, 2001.

[3] T. Li, S. Feng, and L. X. Li, "Information Visualization for Intelligent Decision Support Systems," Knowledge-Based Systems, Vol.14, pp.259-262, 2001.

[4] Y. Liu, et al. "Visualizing Document Classification: A Search Aid for the Digital Library," JASIS, Vol.51, No.3, pp.216-227, 2000.

[5] M. Song, "Visualization in Information Retrieval, A Three-Level Analysis," J. of Information Science, Vol.26, No.1, pp.3-19, 2000.

[6] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons, 1997.

[7] E. Gose, R. Johnsonbugh, and S. Jost, *Pattern Recognition and Image Analysis*, Prentice Hall, 1996.

[8] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/Gather: a cluster-based approach to browsing large document collections," ACM SIGIR '92, pp.318-329, 1992.

[9] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," fifth ACM SIGKDD, pp.16-22, 1999.

[10] H. Schutze and C. Silverstein, "Projections for efficient document clustering," 20th ACM SIGIR, Vol.31, Issue SI, pp.74-81, 1997.

[11] M. Sahami, S. Yusufali, and M. Q. W. Baldonaldo, "SONIA: a service for organizing networked information autonomously," Proc. of DL-98, pp.200-209, 1998.

[12] 이태영, "한국어 초록문의 문장과 내용에 관한 연구", 정보관리연구학회지, 제21권, 제1호, pp.1-33, 1990.

[13] R. K. Maloney, "Title versus Title/Abstract Text Searching SDI System," JASIS, Vol.25, No.6, pp.370-373, 1974.

[14] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," Proc. of Information Visualization, pp.51-58, 1995.

[15] M. Bender, et al. "A functional Framework for Web-Based Information Visualization System," IEEE TVCG, Vol.6, No.1, pp.8-23, 2000.

[16] A. Flexer, "On the use of self-organizing maps for clustering and visualization," Intelligent-Data-Analysis, Vol.5, pp.373-384, 2001.

[17] J. A. Wise, "The ecological approach to text visualization," JASIS, Vol.50, No.13, pp.1224-1233, 1999.

[18] T. Honkela, "Comparisons of self-organized word category maps," In Proceedings of WSOM'97, pp.298-303. 1997.

[19] T. Kohonen, S. Kaski, K. Lagus, and T. Honkela, "Very large twolevel SOM for the browsing of newsgroups," Proc. of ICANN96, Vol.1112, pp.269-274. 1996.

[20] X. Lin, D. Soergel, and G. Marchionini, "A self-organizing semantic map for information

retrieval," In Proc. of the Fourteenth ACM SIGIR, pp.262-269, 1991.

[21] H. Ritter and T. Kohonen, "Self-organizing semantic maps," Biological Cybernetics, Vol.61, No.4, pp.241-254, 1989.

[22] W. B. Frakes and R. B. Yates, *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, 1992.

[23] D. W. Oxtoby and H. P. Gillis, *Principles of Modern Chemistry*, Brooks Cole, 2002.

[24] http://www.research.att.com/~lewis

[25] http://www.wipscorp.com/en

저 자 소 개

지 태 창(Tae-Chang Jee) 정회원

- 1997년 2월 : 연세대학교 컴퓨터과학과 (공학사)
- 1999년 2월 : 연세대학교 컴퓨터과학과 (공학석사)
- 2004년 9월 ~ 현재 : 연세대학교 컴퓨터과학과 (공학박사과정)

<관심분야> : 인공지능, 데이터마이닝, 패턴인식

이 현 진(Hyunjin Lee) 정회원

- 1996년 8월 : 순천향대학교 전산학과 (공학사)
- 1998년 8월 : 연세대학교 컴퓨터과학과 (공학석사)
- 2002년 8월 : 연세대학교 컴퓨터과학과 (공학박사)
- 2003년 1월 ~ 현재 : 한국싸이버대학교 컴퓨터정보통신학부 조교수

<관심분야> : 신경회로망, 데이터마이닝, 이러닝

이 일 병(Yillbyung Lee) 정회원

- 1976년 2월 : 연세대학교 전자공학과 (공학사)
- 1980년 5월 : University of Illinois 전산과학과 (공학석사)
- 1985년 2월 : University of Massachusetts 전산정보과학과 (공학박사)
- 1985년 3월 ~ 현재 : 연세대학교 컴퓨터과학과 교수

<관심분야> : 신경회로망, 문서인식, Computer Vision, Data Mining, 필기체 문자 인식, Biometrics