색인어 가중치 부여 방법에 따른 K-Means 문서 클러스터링의 LSI 분석

오 형 진[†] · 고 지 현[††] · 안 동 언[†††] · 박 순 철[†††]

요 약

정보검색 시스템에서 문서 클러스터링 기술은 사용자 질의에 대해 검색된 문서들을 문서간의 유사도를 기반으로 특정 주제에 따라 재배치하여 놓는 기술로써 사용자에게 검색의 편의성을 제공하고, 그 결과들을 시각적으로 보여줄 수 있다. 본 논문에서는 K-Means 알고리즘을 사용하여 문서를 클러스터링하며 문서를 대표하는 색인어에 가중치를 부여하는 기법에 대하여 논한다. 클러스터링 결과를 시각적으로 보여주기 위하여 문서와 클러스터 중심들을 2차원 공간으로 사상하기 위한 Latent Semantic Indexing 접근 방법을 적용하였다. 실험 결과 문서의 색인어에 대한 가중치 부여 방법을 동일하게 하거나 또는 유사한 수식을 적용한 사례보다는 로컬가중치, 글로벌가중치, 정규화 요소를 모두 부여한 사례에서 문서들이 2차원 벡터 공간에서 군집하여 분포하는 클러스터링 효과가 우수하였다. 특히 로컬 가중치와 글로벌 가중치에 logarithm을 적용하였을 때 문서 분포의 군집도는 현저하게 나타남을 알 수 있었다.

Latent Semantic Indexing Analysis of K-Means Document Clustering for Changing Index Terms Weighting

Hyung-Jin Oh[†] · Ji-Hyun Go[††] · Dong-Un An[†††] · Soon-Chul Park[†††]

ABSTRACT

In the information retrieval system, document clustering technique is to provide user convenience and visual effects by rearranging documents according to the specific topics from the retrieved ones. In this paper, we clustered documents using K-Means algorithm and present the effect of index terms weighting scheme on the document clustering. To verify the experiment, we applied Latent Semantic Indexing approach to illustrate the clustering results and analyzed the clustering results in 2-dimensional space. Experimental results showed that in case of applying local weighting, global weighting and normalization factor, the density of clustering is higher than those of similar or same weighting schemes in 2-dimensional space. Especially, the logarithm of local and global weighting is noticeable.

키워드 : 정보검색, 문서 클러스터링, K-Means 알고리즘, Latent Semantic Indexing

1. 서 론

일반적인 정보검색 엔진에서는 사용자의 질의에 대한 검색 결과를 질의와 문서의 관련도에 따라 매우 긴 목록의 형태로 사용자에게 제공한다. 그러나 오늘날 웹문서의 양이 급격히 증가하고 있으며 작성된 문서의 형태도 다양하기 때문에 일반적인 정보검색 엔진에서는 사용자의 질의를 만족하는 문서를 획득하기 어렵다. 또한 각 검색 엔진마다 검색 결과에 순위를 부여하는 기술을 사용하고 있지만 적합한 문서를 찾아내는 것은 사용자의 몫이다. 따라서 사용자의 요구에 적합한 검색 결과를 선별하여 가공한 후 유용한 지식을 획득하는 문서 클러스터링 기법이 문제 해결 방법으로 등장하였다. 문서 클러스터링은 다량의 문서를 유사한 문서끼리 그룹화하여 특정 주제에 따라 자동 분류하는 것으로써 문서 분류나 데이터 마이닝 분야에서 많이 이용하고 있다[1-3]. 문서 클러스터링 기술을 사용하면 사용자가 특정 정보에 대한 검색을 요구하였을 때 모든 문서를 검색하는 대신 사용자의 요구와 가장 가까운 주제의 클러스터 내의 문서만을 검색함으로써 탐색 시간을 절약할 수 있고 검색의 효율을 향상시킬 수 있다.

본 논문에서는 클러스터링의 성능을 높이기 위하여 클러스터링에 영향을 미치는 요소들 중 각 문서의 색인어 부여 기법을 이용하여 가중치 변화에 따라 클러스터링된 결과를 분석한다[4-6]. 논문의 구성은 다음과 같다. 2장에서는 클러스터링 기법과 관련연구를 살펴보고, 본 논문의 클러스터링 기법으로 구현한 K-Means 알고리즘을 설명한다. 3장에서는

※ 본 연구는 한국과학재단 목적기초연구(R01-2003-000-11588-0) 지원으로 수행되었음.
† 정 회 원 : 3SOFT Technical Consultant
†† 정 회 원 : 삼성테크윈 DSC 개발 센터 요소기술개발 Uni
††† 종신회원 : 전북대학교 전자정보공학부 교수
논문접수 : 2003년 4월 24일, 심사완료 : 2003년 11월 3일

클러스터링 결과 분석을 위해 사용한 Latent Semantic Indexing(LSI)에 대해 알아보고 4장에서는 가중치 부여 기법에 따른 클러스터링 결과를 비교하교 분석한다. 마지막으로 5장에서는 결론을 맺는다.

2. K-Means 알고리즘을 이용한 문서 클러스터링

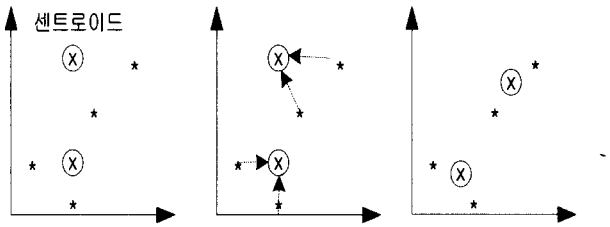
문서 클러스터링은 정보검색의 효율성과 유효성을 증대시키기 위한 목적으로 사용한다. 대표적인 문서 클러스터링 방법론은 클러스터링의 결과로 생성되는 그룹의 구조에 따라서 계층적 클러스터링(hierarchical clustering method)과 비계층적 클러스터링(non-hierarchical clustering method)으로 나눌 수 있는데 각각의 방법론에 따라 여러 가지 구현 알고리즘이 있다.

비계층적 클러스터링은 입력되는 문서의 순서에 따라 클러스터링 결과가 달라지는 단일 처리 방법(single pass method)과 이의 단점을 보완한 재배치 방법(reallocation method)이 있다. 계층적 클러스터링은 문서간의 유사도 정보를 토대로 단계적으로 계층적인 클러스터를 형성하는 방법으로 응집 알고리즘(agglomerative method)과 분할 알고리즘(divisive method)이 있다. 계층적 응집 알고리즘에는 단일 링크 방법(single link method), 완전 링크 방법(complete link method), 그룹 평균 연결 방법(group average link method)등이 있다[7].

일반적으로 대규모 웹문서를 처리하는 클러스터링 알고리즘은 수백만 건을 처리하는 정보검색 시스템 서버에 과부하를 주는 것을 피하고 메모리를 적게 사용해야 할 필요가 있다. 인공지능 분야에서 개발된 기존의 클러스터링 알고리즘들은 고차원의 대규모 데이터 집합으로 문서들이 매우 고차원적이며 sparse한 특성을 나타내어 많은 메모리를 요구한다. 따라서 문서에서 중요한 내용은 포함하면서 중요하지 않은 부분을 제외시킬 수 있다면 클러스터링의 성능에 크게 영향을 미치지 않으면서 메모리를 요구사항을 줄일 수 있다는 장점을 가질 수 있다[6, 8, 9].

정보검색 시스템과 같은 대규모 웹문서를 처리하기 위해서는 클러스터링 속도가 빠르며 구현과 계산 복잡도가 작은 알고리즘을 사용하여 사용자의 검색 요구에 빠르게 응답해야 할 필요성이 있다. K-Means 알고리즘은 비계층적이며 재배치 기법을 사용하는 방법으로써 다양한 어플리케이션에서 사용되는 표준 기술이며, 클러스터링 속도 측면에서 계층적 알고리즘보다 적은 시간이 걸리기 때문에 본 논문에서는 K-Means 알고리즘을 클러스터링 기법으로 사용하였다[10, 11].

K-Means 알고리즘은 비계층적 클러스터링 기법으로 문서와 클러스터의 중심을 나타내는 센트로이드(Centroid)와의 유사도를 측정하여 문서를 적합한 클러스터에 재배치하는 기법이다[12].



(그림 1) 클러스터 중심 생성 과정

여기에서 센트로이드는 클러스터의 중심으로 클러스터에 속하는 문서들의 평균 벡터값을 이용한다. 초기의 클러스터를 형성하고 (그림 1)과 같이 이를 계속적으로 정련하는 과정을 통해 최종의 클러스터를 형성한다. K-Means 알고리즘은 특성상 생성된 클러스터 중심에 따라 클러스터링 결과가 달라진다. 특히 초기 클러스터 중심을 어떻게 선택하는가에 따라 빠른 시간에 최적의 클러스터링 결과가 나오는 경우와 그렇지 않은 경우가 존재한다.

1. K값 클러스터 개수를 구한다.
2. K개의 초기 클러스터 중심(centroid)을 구한다.
3. 각 문서(d)들과 중심값(c) 사이의 거리를 구한다.

$$dist(\overline{d_i}, \overline{c_j}) = \sqrt{\sum_{k=1}^{n}(d_{ki} - c_{kj})^2},$$

$i = 1,2,\cdots n$ n: 전체문서개수
$j = 1,2,\cdots K$ k: *centroid*의개수

4. 문서를 가장 짧은 거리의 중심값에 할당한다.

$$\arg\min dist(\overline{d_i}, \overline{c_j})$$

$$i = 1,\cdots n\ ,\ j = 1,\cdots k$$

$$d_i \in G_c, \quad \text{if } dist(\overline{d_i}, \overline{c_j}) < dist(\overline{d_i}, \overline{c_l})$$

$$(\text{for all } l = 1,2,\cdots,k\ l \neq j)$$

5. 새로운 클러스터 중심값을 다시 계산한다.

$$\overline{c_j} = \frac{1}{|\overline{c_j}|}\sum_{l=1}^{|c_j|}\overline{d_l}$$

6. 이전의 중심값과 새로운 중심값을 비교하여 차이가 거의 없을 때까지 반복한다.

If $\max\ \delta(\overline{c_j^{old}}, \overline{c_j^{new}}) < \theta$ *then return*
식 (7)
else goto 3

(그림 2) K-Means 알고리즘

3. Latent Semantic Indexing 모델

Latent Semantic Indexing(LSI)은 문서의 내용 표현을 서술된 색인어보다는 문서 안에 표현된 개념에 기반을 둔다는 점에 착안하여 제안된 모델이다. LSI는 문서들이 같은 색인어로 구성되어 있지 않더라도 연관성을 나타낼 수 있다. 즉, 어떤 문서가 다른 문서와 개념을 공유한다면 서로

유사한 문서라고 할 수 있다. LSI 모델은 문서들을 저차원 벡터 공간으로 사상시키는 것이다[5, 8, 13].

행렬 m×n으로 나타내는 전체 문서 집합 A는 각 원소, 즉 문서들이 가중치로 표현되는 색인어들을 갖는다. 이때 A를 식 (1)을 이용하여 Singular Value Decomposition(SVD)으로 분해한다.

$$A = U \sum V^T \tag{1}$$

여기에서 U는 단어간 상관 행렬(association matrix)로부터 얻은 m×m 고유 벡터 행렬(orthogonal matrix)이고, V는 문서 간 상관 행렬로부터 얻는 n×n 고유 벡터 행렬이다. 또한 $\sum$는 단일 값을 갖는 m×n 대각 행렬(diagonal matrix)이다. U를 이용하여 단어들은 m차원으로, V를 이용하여 문서들은 n차원으로 사상시킬 수 있다. 동일한 차원으로 단어 벡터와 문서 벡터를 사상시킨다면 단어와 단어의 관계, 단어와 문서간의 관계, 문서와 문서와의 관계를 알 수 있다.

본 논문에서는 가중치 변화에 따른 클러스터링 결과를 분석하기 위한 단계에 LSI 모델을 적용하였다. 각 문서의 색인어에 따른 가중치로 구성된 부분과 클러스터링 결과로 나타난 클러스터 중심값 부분을 행렬식으로 표현한다.

4. 실험 및 클러스터링 결과 분석

4.1 색인어의 가중치 계산 방법

용어 가중치 부여는 문서와 문서를 비교하기 위해서 분류자질 즉 단어에 적절한 가중치를 부여하는 방법이다. 모든 문서에 나타나는 단어는 문서를 구분할 수 있는 색인에서 제외시키는데 이는 사용자가 관심 있는 문서를 구분하는데 영향을 미치지 못하기 때문이다. 반면에 소수의 몇 개의 문서에만 출현하는 단어는 사용자가 관심 있어 하는 문서들을 추출하는데 유용하다. 따라서 문서 내용을 설명하는데 같이 사용된 단어라 할지라도 다양한 비중을 가지고 있으며, 단어의 문서 내에서의 중요성에 대한 척도로서 문서의 각 단어에 가중치를 부여한다[1, 3, 5].

본 논문에서는 m개의 색인어와 n개의 문서로 구성된 문서 집합은 m×n 행렬로서 표현하고 이를 A라고 정의하며 식 (2)와 같다.

$$A = (a_{ij}) \tag{2}$$

여기에서 행렬 A의 각 원소 a_{ij}는 j번째 문서에서 i번째 색인어의 가중치를 의미한다. 행렬의 각 원소 a_{ij}는 식 (3)과 같이 정의한다.

$$a_{ij} = l_{ij} g_i d_j \tag{3}$$

l_{ij}는 j번째 문서에서 i번째 색인어의 로컬 가중치(local weight)이고, g_i는 전체 문서 집합에서 i번째 색인어의 글로벌 가중치(global weight), d_j는 문서의 정규화(normalization) 요소이다.

<표 1>은 로컬 가중치를 부여하는 수식들을 도표화한 것으로서 각 문서의 색인어에 가중치를 부여하여 문서에서 색인어의 중요도를 나타낸다.

〈표 1〉 로컬 가중치 계산 방법

약 어	이 름	수 식
b	Binary	$x(f_{ij})$
l	Logarithmic	$\log(1+f_{ij})$
n	Augmented Normalized Term Frequency	$(x(f_{ij})+(f_{ij}/\max_k f_{kj}))/2)$
t	Term Frequency	f_{ij}

단어의 가중치 계산에 주로 반영하는 요소는 어휘 빈도수, 역문서 빈도수, 문서 길이에 대한 정규화 요소이다. <표 1>에서 Binary는 문서 내에서 색인어의 존재 여부에 따라 값을 부여하는 방법으로 존재하면 1, 존재하지 않으면 0을 부여한다. Binary의 정의는 식 (4)와 같다.

$$x(r) = \begin{pmatrix} 1, & \text{if } r > 0 \\ 0, & \text{if } r = 0 \end{pmatrix} \tag{4}$$

Term Frequency(TF)는 문서 내에서 색인어의 출현 빈도수를 나타내며, Logarithmic 로컬 가중치는 TF가 1인 색인어의 지나치게 낮은 영향력을 보충하고 TF가 높은 단어의 지나친 영향력을 낮추기 위해 TREC-1에서 SMART팀이 제안한 공식이다. Augmented Normalized Term Frequency(보정 로컬 가중치 기법)는 가중치를 일정 범위로 한정시켜서 최소 빈도의 색인어라도 일정 값 이상이 되도록 하면서 동시에 최대 값도 제한하는 공식이다.

〈표 2〉 글로벌 가중치 계산 방법

약 어	이 름	수 식
x	None	1
e	Entropy	$1+\sum_j p_{ij}\log(p_{ij})/\log n)$
f	Inverse Document Frequency(IDF)	$\log(n/\sum_j x(f_{ij}))$
g	GfIdf	$(\sum_j f_{ij})/\sum_j x(f_{ij}))$
n	Normal	$1/\sqrt{\sum_j f_{ij}^2}$
p	Probabilistic Inverse	$\log((n-\sum_j x(f_{ij}))/\sum_j x(f_{ij}))$

<표 2>는 글로벌 가중치(global weight)를 구하는 수식이다. 글로벌 가중치는 전체 문서 집합에서 색인어의 가중치를 표현하는 방법으로 전체 문서를 기준으로 한다는 점에서 로컬 가중치 부여방법과 다르다.

Entropy 글로벌 가중치에서 p_{ij}는 $\frac{f_{ij}}{\sum f_{ij}}$이다. Inverse Document Frequency(IDF)는 역문헌 빈도수로 전체 문서에서 색인어의 빈도수(df : Document Frequency)의 역수로 표현된다. 여기에서 df는 $\sum_j f_{ij}$로 나타낸다. IDF는 가장 많이 사용되고 있는 글로벌 용어 가중치 방법 중의 하나로 $\max(M, \log\frac{n}{df})$로 나타내기도 한다. GfIDf(Global frequency IDF)는 전체 문서에서의 색인어의 빈도수에서 색인어의 출현 빈도수의 비로 나타낸다. 정규 글로벌 가중치 부여 방법은 모든 문서에 대해서 출현 빈도수 제곱의 합의 제곱근의 역수이다. 역 확률 글로벌 가중치는 전체 문서에서의 색인어의 빈도수를 이용하여 확률적으로 나타낸다.

정보검색시스템에서는 길이가 긴 문서일수록 각 단어의 출현빈도가 높고 출현하는 단어의 종류가 많다는 두 가지 원인 때문에 짧은 문헌에 비해서 검색될 확률이 높다는 문제가 발생한다. 이는 클러스터링에서도 마찬가지여서 길이가 긴 문서일수록 다른 문서와의 유사도가 상대적으로 높게 될 여지가 있으므로 문서 정규화가 필요하다.

<표 3>은 문서 정규화를 나타낸다. 문서 정규화는 문서들의 길이를 조절하는 방법으로 보통 각 문서들의 벡터길이를 1로 정규화 한다.

〈표 3〉 문서 정규화의 표현

약 어	이 름	수 식
x	None	1
c	Cosine	$\sum_i (g_i l_{ij})^2)^{-\frac{1}{2}}$

Cosine 정규화는 SMART 시스템에서 사용되는 방법으로서 로컬 용어 가중치와 글로벌 용어 가중치의 조합인 gl을 해당 문서 내 모든 단어의 gl의 제곱의 합의 제곱근으로 나눈다. 예를 들어 가중치를 gfc로 부여한다면 로컬 용어 가중치는 Logarithmic, 글로벌 용어 가중치로는 IDF, 문서 정규화 공식은 Cosine을 이용하는 것이며, 따라서 용어 가중치 부여값은

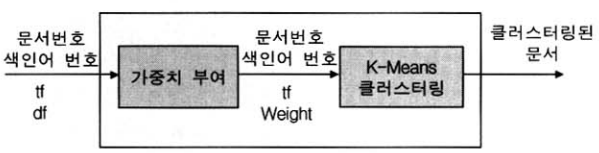
$$a_{ij} = \frac{\log(f_{ij}+1)\log(n/\sum_j x(f_{ij}))}{\sqrt{\sum_i(\log(f_{ij}+1)\log(n/\sum_j x(f_{ij}))^2}}$$

이 된다.

본 논문에서는 로컬 가중치(local weight)와 글로벌 가중치(global weight), 문서 정규화(normalization)를 조합하여 가중치를 계산하였다. 특히 기존의 가중치 부여에서 응용된 복잡한 수식에 의한 가중치 부여를 위하여 로컬 가중치의 단어 빈도수(t)의 복잡도를 $(\frac{t}{t+2})$로 높여서 실험하였다. 이는 단어 가중치에 영향을 미치는 요소 중에서 로컬 가중치의 영향력을 줄이고 그만큼 글로벌 가중치에 비중을 더 주어 글로벌 가중치가 문서-클러스터간 유사도 계산에 기여하는 비중을 늘리기 위해서이다.

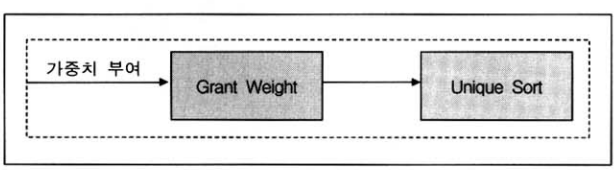
4.2 실험 방법

최근의 정보검색 시스템은 시스템 부하와 메모리를 요구사항, 또한 실시간 처리를 고려하여 검색된 원문 전체를 사용하지 않고 자동으로 문서를 요약한 후 압축된 문서를 사용하는 경향이다. 요약문서 작성은 전체 문서에서 문장별로 중요도를 계산하여 중요도가 높은 문장들을 뽑아내는 방법을 이용하였다. 실험 데이터는 TREC9 데이터 중 102개의 샘플문서를 추출하였다. 본 논문에서는 각 문서 당 색인어, 색인어의 단어 빈도수(Term Frequency), 전체 문서에서 단어 빈도수(Document Frequency)를 이용하여 각 문서마다 색인어별로 중요도를 부여하였다. 본 논문은 정보검색 시스템 Condor Project의 결과물로서 실험데이터 TREC 9 데이터를 사용하여 문서 요약, 문서 클러스터링 기술의 성능을 테스트 하였다. TREC 9데이터는 총 5개의 CD로 구성되어 있으며 각 CD는 24개의 파일로 구성되어 있다. 본 논문에서 실험데이터는 이중 첫 번째 CD의 2번째 파일에서 문서번호 <DOCNO> WTX001-B02-1 </DOCNO>부터 <DOCNO> WTX001-B02-102 </DOCNO>까지 파일크기의 1/4인 102개의 데이터를 선택하여 실험데이터로 사용하였다.



(그림 3) 클러스터링시스템

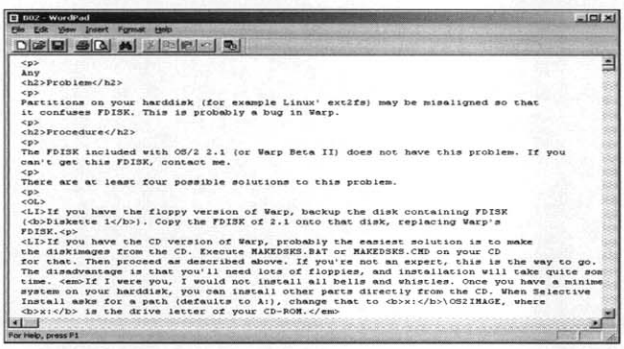
전체 시스템은 문서정보(문서번호, 색인어 번호, tf, df)들을 입력하면 tf와 df를 이용하여 가중치를 부여한다. 또한 tf를 색인어와 문서간의 행렬에서 행(row)의 하나의 위치로 간주하고 K-Means 알고리즘을 이용하여 클러스터링한다. 또한 클러스터링 결과를 분석하기 위한 단계에 LSI 모델을 적용하기 위하여 결과로 나타난 클러스터 중심값을 행렬식으로 표현하였다.



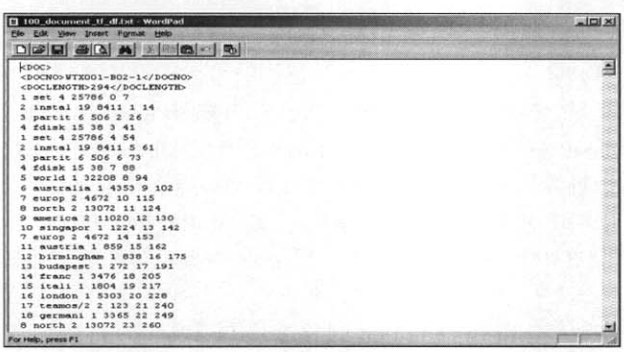
(그림 4) 가중치 부여모듈

가중치 부여 부분은 가중치를 부여하는 부분과 색인어에 따라서 정렬해 주는 부분으로 나뉜다. 색인어에 따른 정렬 시 반복되는 색인어는 제외시켜 준다. 가중치 부여 방법에서는 앞에서 기술한 다양한 가중치 기법을 적용하였다.

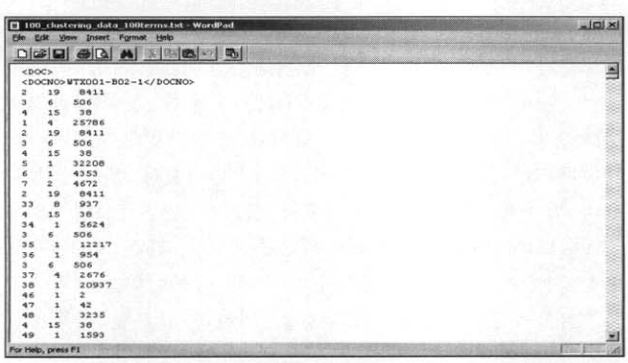
실험 문서 집합의 총 색인어 수는 7,507개이며 102개의 문서이다. 클러스터의 개수(K)는 전체 문서 수에 square root를 사용하여 10개의 클러스터를 생성하였다. 따라서 전체 문헌 집합은 7,507×112 행렬식으로 표현되며 SVD 방법을 이용하여 분해하였다. 실험결과를 표현하기 위하여 문서들 사이의 관계를 V부분만을 이용하였으며 이차원 벡터 공간에 사상시키기 위하여 행렬 V부분을 나타내는 첫 번째와 두 번째 열을 선택하여 x축과 y축에 나타내었다.



(그림 5) 원문



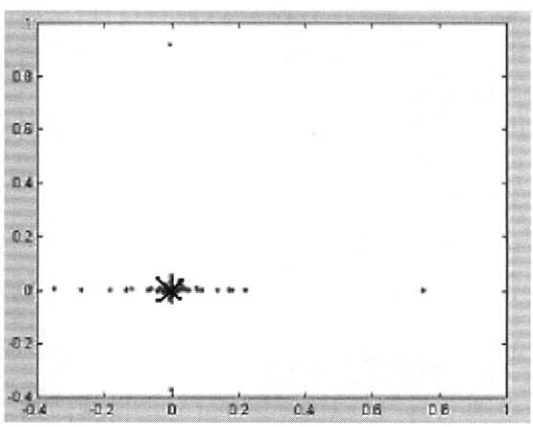
(그림 6) 색인어 데이터



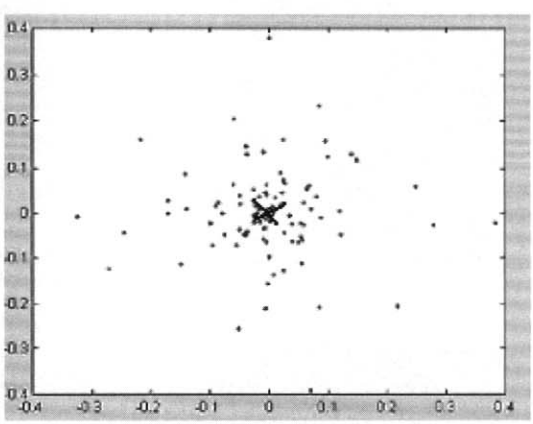
(그림 7) 색인어 번호 부여 후 데이터

4.3 클러스터링 결과 분석

다음 그림들은 3장에서 기술한 가중치 변화에 따른 클러스터링 결과를 LSI 모델을 이용하여 표현한 것이다. "•"으로 표시된 부분은 2차원 벡터공간에서 문서들의 위치를 나타내고, "x"로 표시된 부분은 클러스터 중심의 벡터 공간에서의 위치를 나타낸다. 실험 문서 집합을 나타내는 A의 행렬 표현은 7,507×112이며 SVD 기법을 이용하여 A를 분해한 결과 중에서 문서-문서간의 관련도를 나타내는 벡터인 S는 112×112 행렬이 된다. 유사한 문서일수록 벡터값이 인접한 위치에 표현되기 때문에 동일한 구역에서 점들이 분포함을 발견하였다.



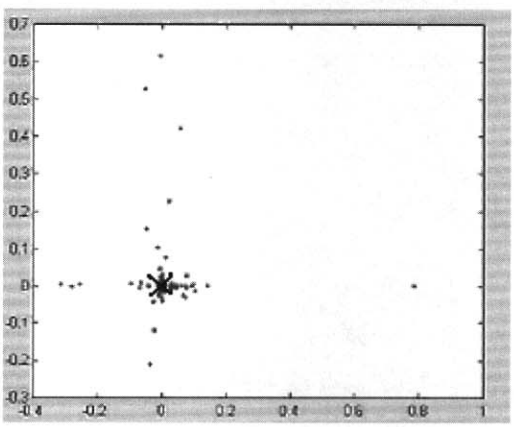
(그림 8) Weight = b×f×x



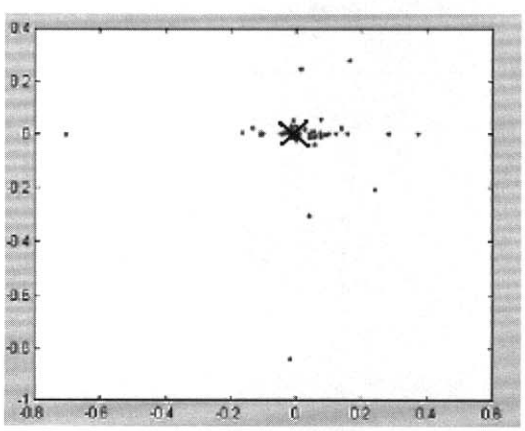
(그림 9) Weight = b×x×x

(그림 8)은 문서의 색인어에 부여하는 가중치 기법 중 로컬 가중치를 Binary를 이용하였고 글로벌 가중치와 문서 정규화는 각각 1로 한 예이다. (그림 9)는 (그림 8)에서 글로벌 가중치 부분만을 역문서 빈도수(Inverse Document Frequency)를 이용하였다. (그림 8)과 (그림 9)에서 색인어 존재 여부만을 가지고 가중치를 부여한 (그림 8)의 경우 문서들의 위치가 (그림 9)에 비해 산재되어 있음을 발견할 수 있다. 이는 각각의 색인어별 가중치가 문서나 전체 문서에서의

중요도에 따른 고유값을 가지지 않기 때문에 문서의 유사도 측정이 어렵다. (그림 8)과 (그림 9)에서 클러스터의 중심 값 위치가 모두 유사한 곳에 있었지만 역문서 빈도수를 적용한 (그림 9)의 경우 동일한 지역에 분포하는 점들이 많았다. 이것은 글로벌 가중치 부여 기법이 클러스터링 결과에 매우 큰 영향을 미친다는 것을 보여준다. 글로벌 가중치는 빈번한 출현 빈도를 가지는 색인어의 가중치 값을 낮춰주는 역할을 하는데 이것은 상대적으로 문서를 대표할 수 있는 색인어의 가중치 값을 높여주게 된다. 따라서 (그림 8)보다는 (그림 9)에서 클러스터링의 효과가 높았다.

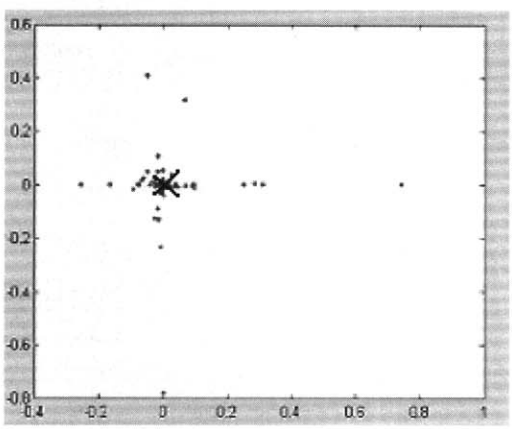


(그림 10) Weight = t / t(t+2)×f×x



(그림 11) Weight = l×f×x

(그림 10)은 로컬 가중치로 문서 내 단어 빈도수 부여방법을 응용한 $\frac{t}{t+2}$(오카피 단어 빈도수)를 이용하여 복잡도를 약간 높인 경우이며, (그림 11)은 단어 빈도수에 logarithm을 적용한 Logarithmic 로컬 가중치를 이용한 예이다. 여기에서 글로벌 가중치와 문서 정규화 요소는 (그림 10)과 (그림 11)에 동일하게 적용하였는데 각각 역문서 빈도수와 1을 선택하였다. (그림 10)과 (그림 11)은 문서들이 한곳에 집중적으로 분포하고 있어서 (그림 8)과 (그림 9)보다 클러스터링이 잘 되었음을 알 수 있다. (그림 9)의 경우 분포의 집중도는 높지만 로컬 가중치 값이 0이 아니면 1인 값을 가지기 때문에 값의 변화가 적어 x축으로만 길게 분포된 반면, (그림 10)과 (그림 11)에서는 x축과 y축으로 문서들이 고른 분포를 보이고 있다. (그림 10)과 (그림 11)의 전체적인 분포도는 비슷한 양상을 보이고 있지만 중심 부분의 점들의 분포를 비교해 보면 로컬 가중치로 오가피 단어 빈도수를 이용한 (그림 10)보다는 logarithm을 적용하여 가중치의 복잡도를 높여준 (그림 11)에서 유사한 벡터값을 가지는 문서가 더 많이 분포함을 발견할 수 있다. 이 경우 (그림 11)이 클러스터링이 더욱 잘 되고 있음을 보여준다.

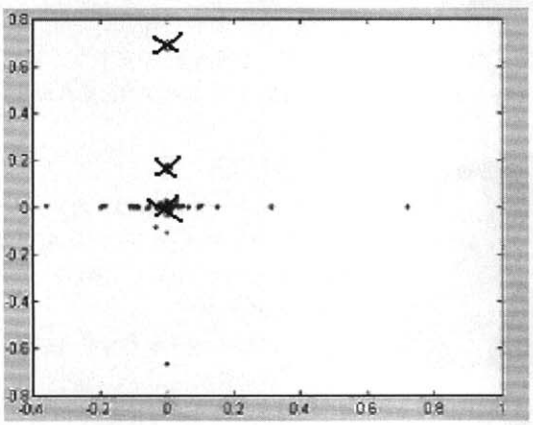


(그림 12) Weight = t/(t+2)×p×x

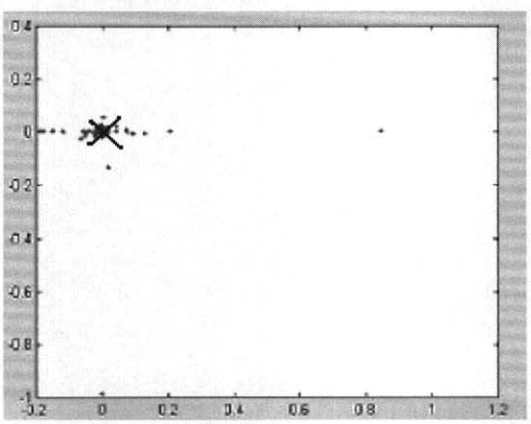
(그림 12)는 (그림 10)에서 글로벌 가중치 부분만을 다르게 한 예이다. 역문서 빈도수대신에 역확률(Probabilistic Inverse)을 사용하였다. (그림 12)는 문서 몇 개를 제외하고는 (그림 10)과 비슷한 분포를 보이고 있다. (그림 10)의 분포에서 0 이상의 x축 위의 점들을 몇 개 제외하면 (그림 12)의 결과와 유사하다. 이것은 가중치 부여에 영향을 미치는 요소로서 역문서 빈도수와 역확률의 차이가 크지 않음을 나타낸다. 특히 logarithm 내의 수식(여기에서는 분자)은 많은 영향을 받지 않는다는 것을 알 수 있다.

(그림 13)과 (그림 14)는 로컬 가중치로 단어 빈도수를 부여한 경우이다. (그림 14)의 경우 글로벌 가중치에 역문서 빈도수를 적용하였다. (그림 13)은 가장 많이 사용하는 가중치 부여 방법인 단어 빈도수(t)를 이용한 것이다. (그림 14)와 비슷한 양상을 보이고 있지만 중심값들의 위치가 크게 차이가 나는 것을 알 수 있다. (그림 14)의 경우 중심값들의 위치가 한 곳에 집중적으로 분포한 반면 (그림 13)은 (그림 14)에 비하여 산재되어있음을 알 수 있다. 이것은 글로벌 가중치인 역문서 빈도수의 영향 때문이다. (그림 13)의 상단 부분에 나타난 중심값의 위치는 글로벌 가중치의 영향뿐만 아니라 이차원 평면으로 사상을 위한 벡터값의 축소로 인한 결과 때문이다. (그림 14)가 (그림 9)와 비슷한

양상을 보이고 있는데 이것 또한 벡터값의 축소로 나타나는 현상이기 때문이다. 그러나 단순 이진 기법을 이용한 (그림 9)보다는 문헌 빈도수를 이용한 (그림 14)의 경우에 클러스터링 결과가 더욱 좋았다. 이것은 색인어에 고유한 값을 부여할수록 각각 고유한 벡터값을 가지기 때문이다.



(그림 13) Weight = t×x×x



(그림 14) Weight = t×f×x

5. 결 론

본 논문에서는 문서 클러스터링에서 문서의 색인어에 부여하는 가중치 부여 기법이 클러스터링 결과에 미치는 영향과 가중치 변화에 따른 클러스터링 결과를 비교 분석하였다. 사용한 클러스터링 기법은 일반적으로 많이 사용하는 비계층적이며 재배치 기법인 K-Means 알고리즘을 이용하였으며 클러스터링 결과를 분석하기 위하여 문서의 색인어에 다양한 가중치를 부여 기법을 적용하였다. 실험 문서 대상 집합과 각 문서는 벡터 공간 표현을 위하여 행렬을 구성하였다. 또한 실험 방법인 가중치 부여 기법에 따른 클러스터링 결과를 2차원 벡터 공간으로 사상시키기 위하여 Latent Semantic Indexing 모델을 사용하였다. 2차원 공간으로 사상하기 위한 벡터 표현 축소로 인하여 문서들의 벡터값이 동일한 예도 있었지만, 상대적으로 유사하지 않은 문서들은 벡터값이 다른 문서와 다른 값을 가지게 되었다. 실험 결과 색인어에 부여한 가중치 부여 방법을 동일하거나 비슷한 수식을 적용한 사례 보다는 색인어에 대한 가중치 부여 방법 중에서 로컬가중치, 글로벌 가중치, 정규화 요소를 모두 부여한 경우 문서들이 2차원 벡터 공간에서 군집하여 분포하는 클러스터링 효과가 우수하였다. 특히 로컬 가중치와 글로벌 가중치에 logarithm을 적용하였을 때 문서 분포의 군집도는 현저하게 나타남을 알 수 있었다. 이것은 문서 내에서 문서를 대표하는 색인어에 가중치를 부여하는 기법이 클러스터링 결과에 매우 영향을 미침을 의미한다. 끝으로 가장 적절한 가중치 부여 기법에 대한 논의와 문서 정규화를 적용한 경우에 클러스터링 결과에 대한 연구가 필요하다.

참 고 문 헌

[1] Ricardo Baeza-Yates, Berhier Ribeiro-NetoRoger, "Modern Information Retrieval," Addison Wesley, 1999.

[2] Michael W. Berry, Murray Browne, "Understanding Search Engines," University of Tennessee, 2001.

[3] 김영택 외 공저, "자연언어처리", 생능출판사, 2001.

[4] Markus Torma, "Comparison Between Three Different Clustering Algorithms," Photogrammetric Journal of Finland, Espoo, Vol.13, No.2, pp.85-95, 1993.

[5] 고지현, 오형진, 박순철, "LSI를 이용한 가중치 변화에 따른 클러스터링결과 분석", 한국정보처리학회, 춘계학술발표논문집, pp.1009-1012, 2002.

[6] 오형진, "클러스터 중심 결정 방법을 개선한 변형 K-Means 알고리즘의 구현", 석사학위 논문, 전북대학교 컴퓨터공학과, 2002.

[7] 이경순, "정보검색에서 벡터공간 검색과 클러스터 분석을 통한 문서 순위 결정 모델", 박사학위 논문, 한국과학기술원, 2001.

[8] 고지현, "정보검색에서 LSI를 이용한 문서 클러스터링에 관한 연구", 석사학위 논문, 전북대학교 정보통신공학과, 2002.

[9] 정영미, 이재윤, "지식 분류의 자동화를 위한 클러스터링 모형 연구", 정보관리학회지, 제18권 제2호, pp.203-230, 2001.

[10] Khaled Alsabti, et al, "An Efficient K-Means Clustering Algorithm," IIPS 11th International Parallel Processing Symposium, 1998.

[11] P. S. Bradley, Uama M Fayyad, "Refining Initial Points for K-Means Clustering," Proceedings of the Fifteenth International Conference on Machine Learning, 1998.

[12] Tapas Kanung, "The Analysis of a Simple k-Means Clustering Algorithm," Proc. of ACM Symposium on Computational Geometry, Hong Kong, June, 2000.

[13] Michael W. Berry, Susan T. Dumais, et al, "Computational Methods for intelligent Information Access," ACM, 1995.

오 형 진

e-mail : hyungjin@3soft.com

1999년 전북대학교 컴퓨터공학과(공학사)

2002년 전북대학교 대학원 컴퓨터공학과 (공학석사)

2001년~2002년 미국 카네기멜론대학 언어기술연구소 방문연구

2003년~현재 3SOFT Technical Consultant

관심분야 : 검색엔진, 문서자동분류, 문서 군집, TDT 등

고 지 현

e-mail : pludy@korea.com

1999년 전북대학교 자연과학대학 수학과 (이학사)

2002년 전북대학교 공과대학 정보통신공학과(공학석사)

2001년~2002년 미국 카네기멜론대학 언어기술연구소 방문연구

2002년~현재 삼성테크윈 DSC 개발 센터 요소기술개발 Uni

관심분야 : 정보검색, 화상처리 알고리즘

안 동 언

e-mail : duan@moak.chonbuk.ac.kr

1981년 한양대학교 전자공학과 (공학사)

1987년 KAIST 전산학과(공학석사)

1995년 KAIST 전산학과(공학박사)

2001년~2002년 전북대학교 정보검색시템 연구센터 센터장

1995년~현재 전북대학교 전자정보공학부 부교수

관심분야 : 정보검색, 한국어정보처리, 문서분류, 문서요약

박 순 철

e-mail : scpark@chonbuk.ac.kr

1979년 인하대학교 응용물리학과(이학사)

1991년 미국 루이지아나주립대학 전산학과 (이학박사)

1991년~1993년 한국전자통신연구원 선임 연구원

1999년~2001년 전북대학교 정보검색시스템센터 센터장

2001년~2002년 미국 카네기멜론대학 언어기술연구소 방문연구

1993년~현재 전북대학교 전자정보공학부 부교수

관심분야 : 정보검색, 영상정보검색