베이지안 SOM과 붓스트랩을 이용한 문서 군집화에 의한 문서 순위조정

최 준 혁† · 전 성 해†† · 이 정 현†††

요 약

기존의 벡터공간 모델을 기반으로 하는 불리언 검색 시스템은 빠른 검색결과를 제공하지만 의미정보를 포함한 사용자의 검색의도를 정확히 반영하지 못한다. 그 결과 사용자 질의어에 대한 검색 결과들은 대부분 사용자 의도와는 전혀 다른 의미의 문서들로서, 사용자들은 이러한 검색문서들 중에서 자신이 원하는 문서를 탐색하기 위해 또 다시 많은 시간을 소요해야 한다. 본 논문에서는 베이지안의 통계적 기법과 무지도 학습의 한 종류인 코호넨 네트워크를 결합한 베이지안 SOM(Self-Organizing feature Maps)을 이용하여, 사용자 질의어와의 의미 유사도에 따라 실시간으로 문서의 군집을 수행한다. 만약, 군집의 대상 문서가 30개미만으로 통계적 특성을 관찰하기 어려운 경우에는 붓스트랩 알고리즘을 이용하여 문서의 개수를 최소 50개 이상으로 확장한다. 이렇게 생성된 군집들은 사용자 질의어와 의미적으로 가장 유사한 문서를 상위에 순위화하기 위하여 각 문서군집의 코호넨 중심값을 이용하여 유사도를 구하며, 이 유사도 값에 의해 2차로 문서순위를 재조정한다.

A Document Ranking Method by Document Clustering Using Bayesian SOM and Bootstrap

Jun Hyeog Choi† · Sung Hae Jun†† · Jung Hyun Lee†††

ABSTRACT

The conventional Boolean retrieval systems based on vector space model can provide the results of retrieval fast, they can't reflect exactly user's retrieval purpose including semantic information. Consequently, the results of retrieval process are very different from those users expected. This fact forces users to waste much time for finding expected documents among retrieved documents. In this paper, we designed a bayesian SOM(Self-Organizing feature Maps) in combination with bayesian statistical method and Kohonen network as a kind of unsupervised learning, then perform classifying documents depending on the semantic similarity to user query in real time. If it is difficult to observe statistical characteristics as there are less than 30 documents for clustering, the number of documents must be increased to at least 50. Also, to give high rank to the documents which is most similar to user query semantically among generalized classifications for generalized clusters, we find the similarity by means of Kohonen centroid of each document classification and adjust the secondary rank depending on the similarity.

1. 서 론

정보검색 시스템의 많은 유형 중에서 군집화를 이용한 정보검색 방법은 검색대상 문서전체를 탐색하는 대신 정보요구 주제와 관련된 문헌군집만을 탐색함으로써 탐색시간의 절약과 검색효율의 향상을 기대할 수 있다. 이러한 관점에서 정보검색 시스템의 검색 결과를 향상시키기 위해 군집화 기법을 이용하는 방법에 대한 연구가 활발히 진행되어 왔다[3, 5, 8].

† 종신회원 : 김포대학 컴퓨터계열 소프트웨어개발 교수
†† 준 회 원 : 인하대학교 대학원 통계학과
††† 종신회원 : 인하대학교 전자계산공학과 교수

논문접수 : 2000년 1월 27일, 심사완료 : 2000년 6월 26일

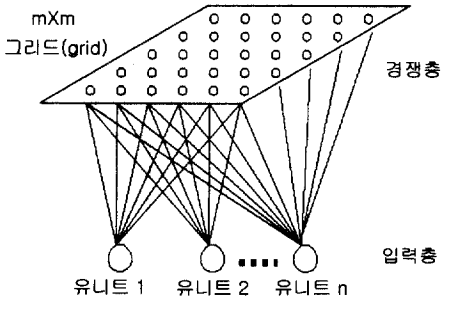
일반적으로 많이 사용하는 문서 군집방법 중에 AHC (Agglomerative Hierarchical Clustering) 알고리즘이 있다. 이 알고리즘은 대상 문서의 수가 많은 경우에 수행 속도가 떨어진다는 단점이 있다. 이를 위해 알고리즘 중지 기준안으로 군집 개수를 사용한다. 이러한 군집방법은 군집속도를 향상시킬 수 있지만, 문서의 군집화가 알고리즘을 중지하기 위한 중지조건에 많은 영향을 받기 때문에 군집화 효율이 떨어지는 결과를 얻을 수 있다. 또 다른 방법으로 Single-link와 Group-average 방법이 있는데, 이는 알고리즘이 수행되기 위해 O(n^2) 시간이 소요되고, Complete-link 방법은 O (n^3)의 시간이 소요된다[5].

본 논문에서는 역화일 중심의 불리언 모델을 이용하는 한국어 정보검색 시스템에서의 정확도 향상을 위하여, 사용자 질의어를 만족하는 검색 결과 문서들을 대상으로 베이지안 SOM을 이용한 실시간 문서 군집방법을 이용한 문서 순위조정 방법을 제안한다. 이를 위해 사용자 질의어와 각 문서의 색인어들과의 엔트로피 값을 계산하여, 이 값을 베이지안 SOM의 입력으로 사용한다.

베이지안 SOM에 의해 군집화된 문서집단은 코호넨 (Kohonen) 중심값을 이용하여 사용자 질의어와의 유사도를 구하며, 이 유사도 값에 따라 군집의 순위를 조정한다.

2. 베이지안 SOM과 문서군집

무감독 학습방법을 사용하고 있는 코호넨 네트워크의 SOM(Self-Organizing feature Maps)은 n-차원의 입력 데이터를 군집화하여, 그 결과를 2차원으로 사상한다.



(그림 1) 코호넨 네트워크 구조

(그림 1)은 일반적인 코호넨 네트워크 구조로 입력 노드들은 모든 출력노드들과 연결되어 가중치값 w_{ij}를 갖는다[1, 2].

(그림 1)의 코호넨 네트워크에서 초기 가중치값은 베이지안의 사전분포를 이용한다. 만약, 학습 데이터에 대한 사전정보를 알고 있으면 이를 초기 연결 가중치 값으로 이용하고, 그렇지 않은 경우에는 주어진 학습 데이터로부터 네트워크 모델의 사전분포를 가장 잘 표현할 수 있는 파라미터를 추정하여 이를 이용한 분포로부터 초기 가중치값을 생성한다. 이러한 사전분포를 이용하면 실제 자료에 대한 정보를 활용한 가중치값을 이용함에 따라 학습시간을 줄일 수 있고 정확한 군집화를 수행하게 된다.

베이지안의 사전분포(prior distribution)는 학습 데이터들로부터 구할 수 있다. 하지만 학습 데이터의 양이 적게 되면 정확한 베이지안의 사전분포를 추정할 수 없게 된다. 이러한 경우 신경망의 학습을 위한 데이터의 양을 충분히 확보하기 위해 통계적 기법인 붓스트랩(Bootstrap) 방법을 사용한다.

붓스트랩 기법은 통계적인 추론에서 사용되며, 분포에 대한 정보없이 제한적으로 주어진 데이터만을 사용하여 파라미터를 추정하는 재표본 추출방법으로 컴퓨터 모의실험을 통해 수행된다. 통계적 관점에서 볼 때 붓스트랩 기법은 데이터만을 이용하여 데이터 분포에 대한 특성을 찾아내는 방법으로 실험에 필요한 많은 양의 데이터를 생성할 수 있는 방법을 제시한다. 따라서 붓스트랩 기법을 이용하여 학습 데이터에 대한 수가 부족할 경우, 부족한 학습 데이터의 양을 보충할 수 있다.

문서의 데이터가 $d_1, d_2, \ldots, d_n$과 같이 n개가 주어졌을 때, SOM의 학습을 위한 데이터 수가 부족할 경우에는 n개의 데이터에서 임의로 한 개를 추출한다. 이렇게 추출된 문서는 다시 원래의 n개 문서집합으로 할당되는 복원추출 방법을 이용한다. N개로 구성된 문서집합에서 또 한 개의 문서를 단순 임의추출하여 조건을 만족할 때까지 반복 수행한다. 이러한 과정을 통해 신경망에서 필요로하는 데이터양은 충분히 확보할 수 있게 된다.

3. 문서군집과 사용자 질의어간의 통계적 유사도

본 논문에서는 불리언 질의를 만족하는 각 문서들의

색인어에 대해 사용자 질의어와의 엔트로피를 구하여, 이 값을 군집화 변수의 값으로 사용하는 문서군집화를 수행한다.

문서의 군집화는 수많은 개개의 문서에 대해서 순위를 부여하는 것보다, 먼저 개개 문서의 색인어에 대해 사용자 질의어와의 엔트로피를 구하여 베이지안 SOM을 이용하여 문서를 집단화한 후에, 이 군집들에 대해 순위를 부여하는 방법이 사용자에게 좀 더 의미있고 정확한 문서들을 상위에 순위화해준다.

실제, N개의 문서들을 p개의 군집변수 각각에 대해 구한 계산결과가 크기 $N\times p$ 자료행렬로 주어졌다고 할 때, 각 문서의 계산값에 대응하는 한 행 벡터(row vector)는 p-차원 공간에서 한 개의 점으로 생각할 수 있다. 이때, p-차원 공간에 N개의 점들이 전체 공간에 걸쳐 임의의 분포로 분산되어 있는지, 혹은 어떤 의미의 친밀성을 가지고 군집을 이루고 있는지에 관한 정보를 갖는다는 것은 사용자 질의어에 의한 문서의 군집화라는 측면에서 중요한 의미를 갖는다.

개개의 문서를 집단화하기 위해서는 우선 문서간의 군집화를 위한 측도가 필요하며, 이를 위해 문서간의 유사성과 비유사성을 이용한다. 본 논문에서 상사나 비상사의 관계는 해당 문서들간의 거리를 통해 요약된다[6].

X_{ik}가 i번째 문서의 k번째 단어와의 엔트로피를 나타내고, $X_i' = (X_{i1}, X_{i2}, \ldots, X_{ip})$은 문서 i의 p개의 엔트로피 값들을 i번째 행벡터로 나타낸다고 가정하자. 그러면 모든 문서는 차원이 $N\times p$인 자료행렬, $X(N\times p)$를 식 (1)과 같이 표현할 수 있다.

$$X_{(N\times p)} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ X_{N1} & X_{N2} & \ldots & X_{Np} \end{bmatrix} = \begin{bmatrix} X_1' \\ X_2' \\ . \\ . \\ X_p' \end{bmatrix} \quad (1)$$

두 문서 X_i'과 X_j'사이의 비상사성을 측정하는 방법은 X_i'과 X_j'간의 거리, $d_{ij} = d(X_i, X_j)$를 계산하여 모든 문서에 대해 식 (2)와 같은 크기 $N\times N$의 거리행렬 D를 얻는 것이다.

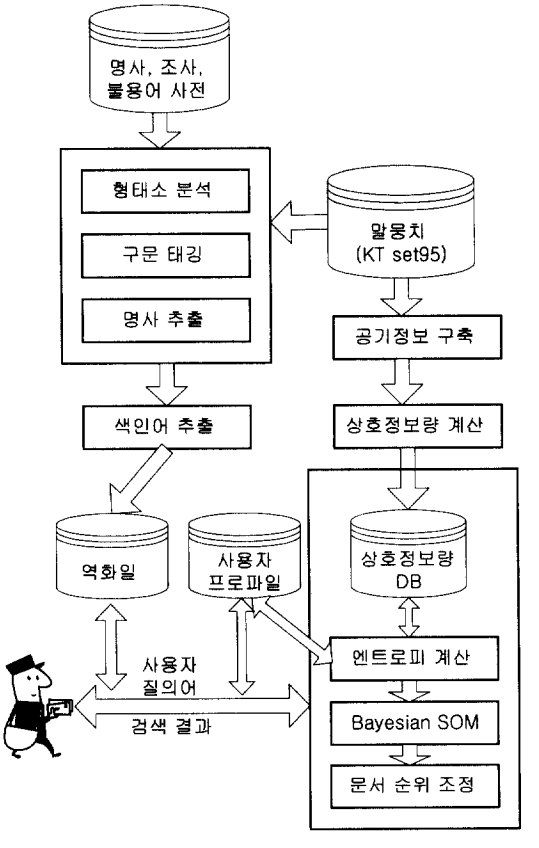
$$D_{(N\times N)} = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1j} & \ldots & d_{1N} \\ d_{21} & d_{22} & \ldots & d_{2j} & \ldots & d_{2N} \\ . & . & \ldots & . & \ldots & . \\ d_{i1} & d_{i2} & \ldots & d_{ij} & \ldots & d_{iN} \\ . & . & \ldots & . & \ldots & . \\ d_{N1} & d_{N2} & \ldots & d_{Nj} & \ldots & d_{NN} \end{bmatrix} \quad (2)$$

본 논문에서의 군집 알고리즘은 식 (2)의 d_{ij}를 원소로 하는 크기 $N\times N$의 거리행렬 D를 사용하여, 상대적으로 거리가 가까운 문서들끼리 같은 군집을 이루게 하여 군집간의 변동에 비해 군집내의 변동을 작게 하는 군집화 방법을 사용한다. 이때, 거리측정을 위한 방법은 많이 있는데, 본 논문에서는 식 (3)과 같은 Minkowski 거리에서 m값이 2인 유클리드 거리를 사용한다[7].

$$d_{ij} = d(X_i, X_j) = \left[\sum_{k=1}^{p} |X_{ik} - X_{jk}|^m \right]^{1/m} \quad (3)$$

4. 전체 시스템 설계

(그림 2)는 엔트로피와 베이지안 SOM을 이용한 문서 순위조정을 위한 전체 시스템 구성도이다.



(그림 2) 베이지안 SOM을 이용한 문서군집 기반의 순위조정 시스템

역화일 구축을 위해 서버에 저장되어 있는 모든 문서들은 형태소 분석을 통한 색인어 추출과정을 거친다. 추출된 색인어 후보들은 Sparck Jones의 역문헌빈도에 의한 가중치[8]에 의해 색인어가 계산되며, 추출된 색인어들은 첫 음절에 의한 동적 트라이 구조로 저장된다[9].

(그림 2)의 전체 시스템을 이용하여 KT set95를 대상으로 사용자 질의어 "한글"을 입력한 결과 문서번호 [0005, 0042, 0052, 0102, 0226, 0311, 0389, 0848, 0919, 4257]의 총 10개의 문서가 적합문서로 검색되었다. (그림 3)은 사용자 질의어 "한글"을 포함하고 있는 문서번호 4257번에 대해, 본 연구실에서 개발한 형태소 분석기를 이용한 형태소 분석 결과이다. 이러한 결과들은 색인어 추출모듈에 의해 각 문서의 색인어 후보에 대해 [단어빈도, 장서빈도, 역문헌빈도에 의한 가중치값, Salton의 가중치값]을 계산하여 백터공간 형태로 저장된다.

<id> 4257
<title> Full-text검색을 지원하는 하이퍼미디어 시스템 개발에 관한 연구
<abstract>본 연구의 최종적인 목적은 Full-text검색을 효율적으로 지원하여 문서관리 및 처리에 적합한 하이퍼미디어 시스템 및 저장시스템의 개발에 있다. [이하 생략]

[검색을 ((검색 을) (N PCO))]
[지원하는 ((지원 하 는)(NH AS ED))]
[하이퍼미디어 ((하이퍼미디어) (N N))]
[시스템 ((시스템) (N))]
[개발에 ((개발 에) (N PCA))]
[관한 ((관하 ㄴ) (V ED)) ((관한) (AD))]
[이하 생략]

(그림 3) 문서번호 4257번에 대한 형태소 분석

<표 1>은 문서번호 4257번을 대상으로 추출한 색인어 후보들에 대한 가중치값이다. 각 문서를 대표하는 색인어 수는 엔트로피 계산에 따른 시간 지연을 고려하여 가중치값의 크기에 따라 3개의 색인어를 결정한다. 따라서 문서번호 4257을 대표하는 색인어는 [하이퍼, 문서, 검색]이 된다.

각 문서와 사용자 질의어와의 엔트로피 계산은 색인어와 사용자 질의어간의 공기정보를 이용한 상호정보량을 이용하여 계산한다.

〈표 1〉 문서번호 4257번에 대한 가중치계산

색인어 후보	용어 빈도	문헌 빈도	Sparck Jones의 가중치	Salton의 가중치
가지	1	606	2.8073	2.8073
개발	2	1939	1.0000	2.0000
검색	5	232	4.2479	21.230
검증	1	110	5.3219	5.3219
관리	2	698	2.584	5.1699
구조	1	708	2.584	2.5849
구축	2	621	2.8073	5.6147
기능	2	869	2.3219	4.6438
..	..	..	..	..

만약, 어떤 문서 d_i의 색인어가 각각 $term_1$, $term_2$, $term_3$라고 하면, 사용자 질의어와 이들 주제어들 사이의 엔트로피는 식 (4)에 의해 구할 수 있다[9].

$$H_{[\text{질의어, Term}ij]} = -\sum_{j=1}^{3} MI_{[\text{질의어, term}ij]} \log_2 MI_{[\text{질의어, term}ij]} \quad (4)$$

식 (4)에서, i는 $1,\cdots,N$으로 N은 문서의 개수를, j는 $1,\cdots,n$으로 n은 색인어의 개수를 나타낸다. 또한, $H_{[\text{질의어, Term}ij]}$는 질의어와 색인어간의 상호정보량을 나타낸다.

<표 2>는 문서번호 4257번에서 식 (4)를 이용하여 계산한 각 문서의 주제어들과 사용자 질의어와의 엔트로피 계산 결과이다. 이러한 엔트로피값은 베이지안 SOM의 입력으로 사용되어 문서군집을 수행한다.

〈표 2〉 사용자 질의어 "한글"과 각 문서의 주제어들과의 엔트로피

ID	문서번호	엔트로피1	엔트로피2	엔트로피3
1	0005	1.99033	1.99612	1.97138
2	0042	1.99434	1.76747	1.78802
3	0052	1.97207	1.96379	1.98480
4	0102	1.98645	1.74576	1.99971
5	0226	1.95489	1.77603	1.98795
6	0311	1.99441	1.76259	1.96714
7	0389	1.98132	1.48992	1.99853
8	0848	1.96501	1.98282	1.98647
9	0919	1.99285	1.95416	1.99505
10	4257	1.98790	1.97150	1.99858

군집화된 문서집단은 사용자 질의어와의 유사도를 계산하기 위하여 각 문서집단의 코호넨 중심값을 이용한다. 이때의 유사도는 식 (5)와 같은 노름(Norm) 값

을 이용한다.

임의의 코호넨 중심 : $C=(c_1,c_2,\ldots,c_n)$

코호넨 중심의 Norm = $\sqrt{c_1^2+c_2^2+\ldots+c_n^2}$ (5)

임의의 코호넨 중심은 학습이 끝난 코호넨 네트워크의 출력층에 대한 최종 가중치값이다. 계산이 끝난 후, 각 문서군집의 코호넨 중심값은 코호넨 중심벡터의 노름값으로 정의하고, 이 값이 가장 큰 값이 사용자가 원하는 문서집단으로서 다른 문서집단들보다 상위로 순위조정된다.

5. 실험 및 평가

본 논문의 실험은 SMP NT 서버 6400Qp상에서 한국통신에서 구축한 정보검색 시험용 데이터 모음인 KT set95의 4,414개의 문서에 대해 수행하였다.

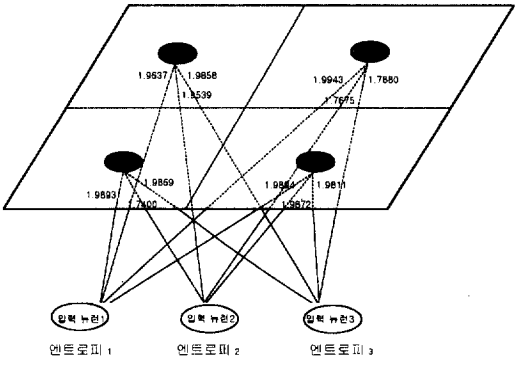
KT 문서집합 자체가 한정된 도메인 상에서 작성된 데이터이기 때문에 사용자 질의어를 이용하여 검색을 수행하면 해당되는 문서는 최대 50개를 넘지 않는다. 따라서, 본 논문에서 설계한 베이지안 SOM을 이용하여 군집화를 수행하기 위해서는 데이터의 수가 부족하다. 이러한 문제를 해결하기 위해 본 논문에서는 붓스트랩 기법을 이용하여 50개미만의 데이터 수를 최소 50개 이상이 되도록 데이터 수를 확장하였다.

〈표 3〉 문서번호 4257번에서 추출한 엔트로피 벡터를 대상으로 붓스트랩 알고리즘을 적용한 결과

ID	문서번호	엔트로피1	엔트로피2	엔트로피3
3	0052	1.97207	1.96379	1.98480
6	0311	1.99441	1.76259	1.96714
2	0042	1.99434	1.76747	1.78802
6	0311	1.99441	1.76259	1.96714
3	0052	1.97207	1.96379	1.98480
4	0102	1.98645	1.74576	1.99971
9	0919	1.99285	1.95416	1.99505
6	0311	1.99441	1.76259	1.96714
6	0311	1.99441	1.76259	1.96714
...	...	...	...	...
10	4257	1.98790	1.97150	1.99858
8	0848	1.96501	1.98282	1.98647

<표 3>은 사용자 질의어 "한글"을 만족하는 적합문서들에서 추출한 사용자 질의어와 색인어들과의 엔트로피 계산결과를 대상으로 붓스트랩 기법을 적용하여 실험대상 문서 수를 50개로 확장한 결과이다. 이 값은 베이지안 SOM의 입력으로 사용된다. 이때, 신경망의 초기 연결 가중치 값은 파라미터 0과 1을 갖는 균일분포(uniform distribution), U(0,1)을 이용하였다.

<표 3>의 50개의 입력 데이터에서 엔트로피 개수가 3개이기 때문에 입력뉴런은 3으로, 출력뉴런은 2×2의 4개의 출력노드 구조를 사용하였으며, 학습율은 1로 하였다. 이는 베이지안 SOM의 사전, 사후분포를 이용한 연결 가중치의 변화가 학습률과 같은 다른 요인에 의해 영향을 받지 않도록 하기 위함이다. 학습율의 변화를 고려하여 실험을 수행할 수 있지만, 본 논문에서는 순수하게 베이지안 SOM의 사전, 사후 분포만을 이용하여 가중치를 갱신하는 방법을 적용하였다.



(그림 4) 사용자 질의어 "한글"에 대한 베이지안 SOM의 학습결과

(그림 4)는 <표 3>을 대상으로 베이지안 SOM을 이용한 군집 결과이다. 각 군집에 대한 문서의 할당결과를 보면 문서군집1에는 문서번호 [0226, 0919]번이 할당되었고, 문서군집2에는 문서번호 [0005, 0052, 0102, 0389, 0848, 4257]번이 할당되었으며, 문서군집3에는 문서번호 [0042, 0311]번이 할당되었다.

(그림 4)에서 알 수 있듯이 각 문서군집의 연결가중치는 각각의 엔트로피에 대해서 문서군집1은 {1.9637, 1.9539, 1.9858}, 문서군집2는 {1.9943, 1.7675, 1.7880}, 문서군집3은 {1.9893, 1.7400, 1.9869}로 결정되었다. 이러한 연결가중치는 질의어에 대한 유사도와 같은 개념으로 최종 연결가중치 벡터의 크기가 큰 문서군집이 사용자 질의에 대한 유사도가 높은 문서군집으로 간주되어 문서군집 순위를 재조정한다.

최종 연결가중치 벡터의 크기는 식 (5)의 노름값을 이용한다. 식 (5)를 이용하여 계산한 각 문서군집의 노

름값은 문서군집1이 3.4105이고, 문서군집2는 3.2091, 문서군집3은 3.3065이다. 따라서 연결가중치 벡터의 노름값이 가장 큰 문서군집1을 가장 상위에 위치시키고, 다음으로 문서군집3, 문서군집2의 순으로 군집의 순위를 조정한다.

엔트로피와 군집 수에 대한 증가는 서버에서의 계산양을 가중시켜 검색지연의 한 요인이 될 수 있다. 따라서, 엔트로피와 문서군집 수의 증가에 따른 베이지안 SOM의 학습시간 증가를 검토해 볼 필요가 있다.

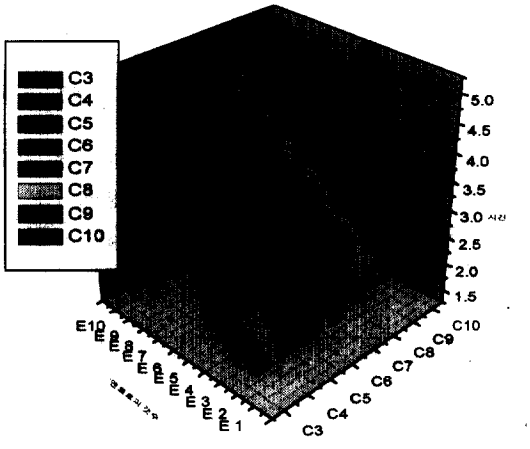
<표 4>는 엔트로피와 군집시간에 대한 상관계수를 나타낸다. 이에 대한 실험환경은 CPU Pentium II 350 MHz, RAM 128M의 윈도우 NT 서버로서 초기 학습을 위한 실험 데이터는 900개의 문서를 사용하였으나 문서군집을 수행하는 속도가 모두 0.01초미만으로 붓스트랩 알고리즘을 이용하여 문서의 수를 최대 149,760개로 확대하여 실험해 보았다. 이는 11.5MB에 이르는 방대한 문서의 크기로 어떤 한 질의어에 의해 검색되는 문서의 개수가 이를 초과할 것으로 생각되지 않는 매우 큰 문서의 양이다.

〈표 4〉 엔트로피와 군집 수에 따른 시간 분석

클러스터 / 엔트로피	C3	C4	C5	C6	C7	C8	C9	C10
E1	1.77	1.80	1.82	1.92	1.97	2.13	2.27	1.73
E2	1.89	2.53	2.14	2.26	2.36	2.48	2.50	2.00
E3	2.06	2.25	2.37	2.48	2.68	2.75	2.75	2.27
E4	2.28	2.43	2.65	2.69	2.78	3.10	3.13	2.63
E5	3.58	2.72	2.95	2.99	3.11	3.35	3.29	2.76
E6	2.64	2.86	2.95	3.30	3.37	3.49	4.26	3.95
E7	2.85	3.02	3.14	3.44	3.58	3.70	4.45	4.04
E8	3.72	3.87	4.15	3.61	3.67	3.92	4.12	4.23
E9	3.88	3.41	3.71	3.83	3.99	4.18	4.43	4.63
E10	3.35	3.61	4.67	4.11	4.33	4.56	4.74	4.95

(그림 5)는 <표 4>를 이용한 엔트로피와 문서군집 수에 따른 군집시간에 대한 상관관계를 나타낸다.

(그림 5)의 군집시간을 살펴보면 엔트로피 수에 따라 학습시간이 증가됨을 알 수 있다. 그러나 엔트로피의 개수를 최대 10, 생성될 수 있는 군집의 개수를 최대 10으로 하더라도 계산을 위해 소요되는 시간은 최대 5초미만이다. 이는 정확도 향상을 위한 사용자의 대기시간으로 허용 범위안에 있는 시간이라 생각된다.



(그림 5) 엔트로피와 문서군집 수에 따른 군집시간

<표 5>와 <표 6>은 사용자 질의어 20개에 대한 역화일 중심의 벡터공간 모델에 대한 베이지안 SOM의 검색효율에 대한 비교표이다.

〈표 5〉 검색효율 비교표 #1

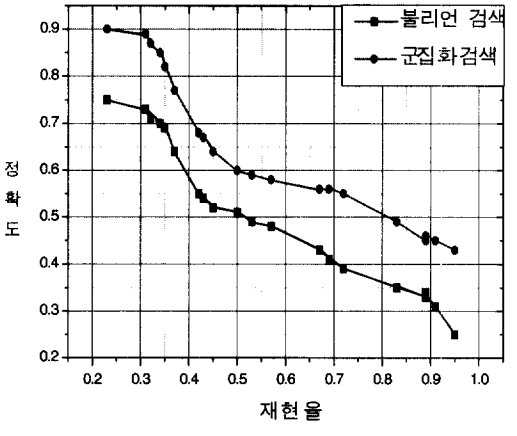
키워드	불리언 검색		군집 검색	
	재현율	정확도	재현율	정확도
정보검색	0.72	0.39	0.72	0.55
전자파장해	0.35	0.69	0.35	0.82
자동차	0.23	0.75	0.23	0.90
펜티엄	0.45	0.52	0.45	0.64
방사능	0.31	0.73	0.31	0.89
레이저	0.57	0.48	0.57	0.58
냉 매	0.67	0.43	0.67	0.56
질 병	0.91	0.31	0.91	0.45
이동통신	0.42	0.55	0.42	0.68
세탁기	0.83	0.35	0.83	0.49

〈표 6〉 검색효율 비교표 #2

키워드	불리언 검색		군집 검색	
	재현율	정확도	재현율	정확도
교환기	0.32	0.71	0.32	0.87
멀티미디어	0.95	0.25	0.95	0.43
병렬컴퓨터	0.50	0.51	0.50	0.6
음성인식	0.53	0.49	0.53	0.59
증권	0.69	0.41	0.69	0.56
형태소	0.89	0.33	0.89	0.45
자동번역	0.37	0.64	0.37	0.77
지리정보	0.87	0.34	0.87	0.46
신용카드	0.34	0.70	0.34	0.85
문자인식	0.43	0.54	0.43	0.67

검색 효율 비교를 위해 본 논문에서 사용하는 재현율-정확도는 정보검색 효율의 척도를 계산하는 일반식을 이용하되[8], 전체 문서의 정확도 중에서 상위 10개 문서에 대한 사용자의 정확도를 계산하였다.

(그림 6)은 <표 5>와 <표 6>을 기반으로 한 정확도-재현율의 관계로 베이지안 SOM을 이용한 군집 검색이 불리언 중심의 역화일 방법에 비해 평균 14.5%의 향상된 정확도를 나타내고 있다.



(그림 6) 정확도-재현율 비교율

6. 결 론

본 논문에서는 자기 조직화 특성이 있는 베이지안 SOM을 이용하여 사용자 질의어와 해당 문서의 의미 유사도에 따른 문서군집 기반의 문서 순위조정 시스템을 설계하고 구현하였다. 이때, 베이지안 SOM의 입력값으로는 사용자 질의어와 각 문서의 주제어에서 추출한 상호정보량에 기반한 엔트로피값을 이용하였다. 만약, 베이지안 SOM의 입력으로 사용되는 문서의 개수가 30개미만으로 통계적 특성을 파악하기 어려운 경우에는 붓스트랩 알고리즘을 이용하여 학습문서의 개수를 50개 이상으로 확장하였다. 그 결과 본 논문에서 설계한 베이지안 SOM은 기존의 역화일 중심의 벡터 공간 모델보다 재현율의 저하없이 평균 14.5%에 대한 정확도 향상을 얻을 수 있었다.

향후 연구 과제로는 언어의 일반 현상을 설명할 수 있는 대규모의 말뭉치를 구축하는데 있다. 이러한 말뭉치를 구축하여 단어간 확률 정보인 엔트로피를 계산한다면 사용자의 의도를 더욱 정확히 반영할 수 있어 실질 정확도는 증가될 것으로 기대한다.

참 고 문 헌

[1] T. Kohonen, Self Organizing Maps, 2nd Edition, Springer, 1997.

[2] T. Kohonen, Self-Organization and Associative Memory, Springer-Verlag, 2nd ed. 1988.

[3] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[4] Johnson, Richard A. and Wichern Dean W, Applied Multivariate Statistical Analysis, Prentice Hall, 1992.

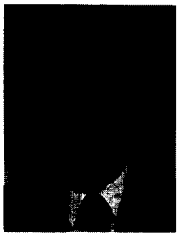
[5] Oren Zamir, Oren Etzioni, "Web Document Clustering : A Feasibility Demonstration," Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.46-54, 1998.

[6] 김기영, 전명식, 다변량 통계자료분석 , 자유아카데미, 1994.

[7] 김기영, 전명식, SAS 군집분석 , 자유아카데미, 1992.

[8] 정영미, 정보검색론, 구미무역, 1993.

[9] 최준혁, 허준회, 이정현, "한국어 정보 검색에서 엔트로피와 사용자 프로파일을 이용한 질의 확장", 한국통신학회논문지, 제24권 제11호, pp.1729-1738, 1999.



최 준 혁

e-mail : jhchoi@kimpo.ac.kr

1990년 경기대학교 전자계산학과 졸업

1995년 인하대학교 대학원 전자계산공학과 졸업(공학석사)

2000년 인하대학교 대학원 전자계산공학과 졸업예정 (공학박사)

1997년~현재 김포대학 컴퓨터계열 (소프트웨어개발 전공) 조교수

관심분야 : 정보검색, 신경망, 데이타 마이닝, 자연어처리

전 성 해

e-mail : shjun@anova.inha.ac.kr

1993년 인하대학교 이과대학 통계학과 졸업(학사)

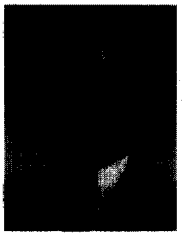
1996년 인하대학교 대학원 통계학과 (이학석사)

1999년 인하대학교 대학원 통계학과 (박사수료)

1996년~1997년 효성그룹 전자통신연구소 연구원

2000년~현재 한서대학교 컴퓨터 통신공학과 겸임교수

관심분야 : 신경망, 데이타 마이닝



이 정 현

e-mail : jhlee@inha.ac.kr

1977년 인하대학교 전자공학과 졸업

1980년 인하대학교 대학원 전자공학과(공학 석사)

1988년 인하대학교 대학원 전자공학과(공학박사)

1979년~1981년 한국전자기술연구소 시스템 연구원

1984년~1989년 경기대학교 전자계산학과 교수

1989년~현재 인하대학교 전자계산공학과 교수

관심분야 : 자연언어처리, HCI, 정보검색, 음성인식, 음성합성, 계산기 구조