2002년 한국정보처리학회 춘계학술발표논문집 제9권 제1호

신경망 GHSOM을 이용한 의료 문헌 정보의 군집화

허진석*, 김인철**
경기대학교 전자계산학과
e-mail:{hjs0823*, kic**}@kyonggi.ac.kr

Medical Document Clustering using the Growing Hierarchical SOM

Jin-Seok Heo*, In-Cheol Kim**
Dept of Computer Science, Kyonggi University

요 약

일반적으로 PubMed와 같은 인터넷을 이용한 대규모 의료 문헌정보 검색시스템에서 포괄적인 주제어나 간결한 주제어를 이용한 검색을 시도할 경우, 종종 매우 다양한 세부주제의 문헌리스트들이 다량으로 검색된다. 이러한 경우 이용자는 실제로 본인이 원했던 세부주제에 부합되는 문헌들을 찾기 위해서는 검색결과로 주어진 긴 문헌리스트상의 문헌 하나하나에 대해 다시 문헌제목이나 혹은 요약 등의 내용을 직접 읽어보고 내용을 확인하여야 한다. 이러한 작업은 매우 번거럽고 시간과 노력을 많이 필요로 한다. 따라서 본 논문에서는 이러한 노력을 줄이기 위한 한 가지 방안으로, PubMed시스템의 주제어 검색결과로 주어진 문헌들에 대해 내용의 유사성과 차별성에 따라 자동으로 몇 개의 그룹으로 나누어주는 군집화시스템 MedCluster의 설계와 구현에 대해 소개한다. MedCluster의 큰 특징은 기존의 문서 군집화 방법과는 다른 신경망 GHSOM을 이용한 군집화 방법을 사용하는 점이다. GHSOM은 미리 문서 그룹의 개수를 정해줄 필요가 없고 다양한 레벨의 문서 그룹들을 얻을 수 있는 계층적 군집화를 이루어낸다는 장점을 가지고 있다. 본 논문에서는 신경망 GHSOM의 구조와 특성에 대해 간략히 살펴보고, GHSOM을 채용한 의료문헌 군집화시스템 MedCluster의 설계와 구현에 대해 설명한다.

1. 서 론

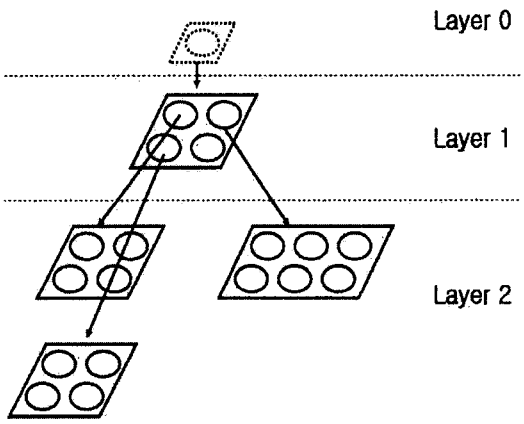
최근들어 인터넷이 발전하면서 유용한 다량의 문헌자료들이 인터넷을 통해 손쉽게 접근할 수 있게 되었다. 특히 의학, 약학, 생물학과 같은 생명과학 연구에 매우 귀중한 문헌정보를 제공해온 대표적인 의료문헌데이터베이스인 MEDLINE도 PubMed와 같은 형태로 인터넷을 통해 다수의 연구자에게 정보를 제공하게 되었다. 하지만 일반적으로 PubMed와 같은 보유 자료가 방대한 대규모 의료 문헌정보 검색 시스템에서는 이용자가 비교적 포괄적인 주제어나 간결한 주제어를 이용한 검색을 시도할 경우, 종종 매우 다양한 세부주제의 문헌리스트들이 다량으로 검색된다. 이러한 경우 이용자는 실제로 본인이 원했던 세부주제에 부합되는 문헌들을 찾기 위해서는 검색결과로 주어진 긴 문헌리스트상의 문헌 하나하나에 대해 다시 문헌제목이나 혹은 요약 등의 내용을 직접 읽어보고 내용을 확인하여야 한다. 이러한 작업은 매우 번거럽고 시간과 노력을 많이 필요로 한다.

본 논문에서는 이러한 노력을 줄이기 위한 한 가지 방안으로, PubMed시스템의 주제어 검색결과로 주어진 문헌들에 대해 내용의 유사성과 차별성에 따라 자동으로 몇 개의 그룹으로 나누어주는 군집화시스템 MedCluster의 설계와 구현에 대해 소개한다. MedCluster의 큰 특징은 기존의 문서 군집화 방법과는 다른 신경망 GHSOM을 이용한 군집화 방법을 사용하는 점이다. GHSOM은 미리 문서 그룹의 개수를 정해줄 필요가 없고 다양한 레벨의 문서 그룹들을 얻을 수 있는 계층적 군집화를 이루어낸다는 장점을 가지고 있다. 본 논문에서는 신경망 GHSOM의 구조와 특성에 대해 간략히 살펴보고, GHSOM을 채용한 의료문헌 군집화시스템 MedCluster의 설계와 구현에 대해 설명한다.

2. 관련 연구

군집화(clustering)는 상호 유사성에 따라 주어진 데이터들을 자동으로 몇 개의 군집으로 나누는 과정을 말하며, 이를 위해 주로 인공지능의 비교사 학습(unsupervised learning) 알고리즘이나 통계적 방법들을 많이 이용한다. 문서 군집화에 많이 이용되는 기존의 방법들은 크게 분할(partitioning) 방법과 계층적(hierarchical) 방법으로 나눌 수 있다. 분할방법은 주어진 문서 집합을 미리 정해진 K개의 분할 영역으로 나누는 방법이다. 대표적인 분할방법들로는 K 개의 중심점과 각 문서모델간의 Euclidean distance에 기초한 K-means와 K-medoid 방법이 있다. 계층적 방법은 개별 문서 하나하나를 하나의 군집으로 설정한 다음, 각 문서 간의 거리를 기초로 가장 가장 가까운 문서들끼리 합병해가는 Bottom-up 또는 Agglomerative 방식과 모든 문서들을하나의 군집으로 설정한 뒤, 그 군집 내에서 이질성이 높은 세부 군집들을 찾아 나누어 가는 Top-Down 또는 Divisive 방식이 있다. 신경망 SOM(Self-Organizing Map)은 분할 군집화 방법의 하나로서, 각 문서 그룹에 대응되는 K개의 뉴런들로 구성된 1차원 혹은 2차원 단일 계층의 신경망이다. SOM에서는 하나의 문서모델이 입력벡터로 주어지면 각 뉴런의 가중치벡터(weight vector)들과 비교하여 입력과 가장 유사한 가중치를 갖는 뉴런에 입력문서가 배정되고, 이 뉴런과 이웃한 뉴런들의 가중치를 입력에 가깝게 조정하는 학습과정이 되풀이 된다. 이와 같은 학습과정을 거쳐 SOM은 입력되는 문서들에 대해 뉴런별로 하나의 군집을 형성하게 되고, 또한 유사한 문서 군집들의 뉴런은 거리적으로도 가까이 배치되어 군집화의 결과를 시각적으로 이해하기 좋은 장점을 가진다. 하지만 SOM을 이용하기 위해서는 미리 적당한 문서 군집수를 예측하고 그 수만큼의 뉴런들로 신경망을 구성해야하는데, 문서들의 내용을 모두 확인하지 않은 상태에서는 이러한 문서 군집수를 정하는 일은 적절하지 못하다. 성장 SOM(Growing SOM, GSOM)은 이러한 SOM의 문제점을 극복하기 위해, 입력되는 문서들의 양과 이질성에 따라 뉴런의 수를 늘여감으로써 스스로 군집의 개수를 정해가는 특징을 가지고 있다. 한편, 계층적 SOM(Hierarchical SOM, HSOM)은 SOM이나 GSOM의 경우와는 달리 주어진 문서들을 단순히 K개의 군집으로 분할하지 않고, 유사성과 크기가 다른 연속된 여러 계층의 군집들을 생성해내는 특징이 있다.

앞서 소개한 문서 군집화 방법을 이용한 다양한 응용 사례 연구들이 있다. 전자도서관 시스템인 헬싱키 대학의 SOMlib, 스탠포드 대학의 SONIA, 텍스트마이닝 상용시스템인 IBM의 Intelligent Miner for Text, Vivisimo 등이 대표적인 예들이다.



(그림 1) GHSOM의 구조

3. GHSOM

GHSOM (Growing Hierarchical SOM) 은 성장 SOM(GSOM)과 계층적 SOM(Hierarchical SOM)의 장점을 결합하여 만들어진 군집화 알고리즘이다. GHSOM의 기본적인 구조는 (그림 1)과 같이 여러 계층의 서로 독립적인 SOM들로 구성된 계층적 구조로 되어있고, 기존 SOM과는 달리 입력 문서들에 따라 맵(map)의 크기와 계층 수가 스스로 늘어나는 성질을 가지고 있다. GHSOM에서 계층은 크게 계층 0, 계층 1, 그리고 나머지 부분으로 구분 할 수 있다. 우선 계층 0은 가상의 계층으로 1개의 유닛(unit)을 포함하고 있으며, 이 유닛의 가중치벡터는 $m_0 = [\mu_{01}, \mu_{02}, \ldots, \mu_{0n}]^T$ 와 같이 표현되며, 모든 입력 데이터의 평균으로 초기화 된다. 입력데이터 $\boldsymbol{x}$ 와 이 유닛과 편차는 (식 1)과 같이 나타낼 수 있다.

$$mqe_0 = \frac{1}{d} \bullet \| m_0 - x \|, \quad d: x\text{의 수} \qquad \text{(식 1)}$$

mqe_0를 계산한 후에 GHSOM의 첫번째 SOM으로부터 훈련이 시작된다. 첫번째 계층의 맵은 유닛의 수보다 적은 수의 유닛으로 초기화된다. 각 유닛 i는

(식 2)과 같이 n-차원의 벡터 m_i로 정의된다.

$$m_i = [\mu_{01}, \mu_{02}, ..., \mu_{0n}]^T, m_i = \Re^n \quad \text{(식 2)}$$

각 유닛들은 랜덤한 값으로 초기화되며, SOM의 학습 규칙은 (식 3)과 같다. 여기서 a는 학습률이고, h_{ci}는 이웃함수(neighborhood function)이고, x는 현재의 입력 패턴이다. 그리고 c는 t 만큼 반복한 후의 승자 유닛이다.

$$m_i(t+1) = m_i(t) + \alpha(t) \bullet h_{ci}(t) \bullet [x(t) - m_i(t)] \quad \text{(식 3)}$$

일정한 회수만큼의 훈련이 이루어진 후에 (식 4)에 의해 맵의 MQE(Mean Quantization Error)가 계산된다. 여기서 u는 SOM m 에 포함된 유닛 i의 갯수이고, mqe_i에 의해서 계산된다.

$$MQE_m = \frac{1}{u} \bullet \sum_i mqe_i \quad \text{(식 4)}$$

그리고 (식 5)의 조건이 만족하는 동안 mqe_e가 가장 큰 유닛 e 에 새로운 열이나 행을 삽입함으로써 맵은 계속 성장한다.

$$MQE_m \geq \tau_m \bullet mqe_0 \quad \text{(식 5)}$$

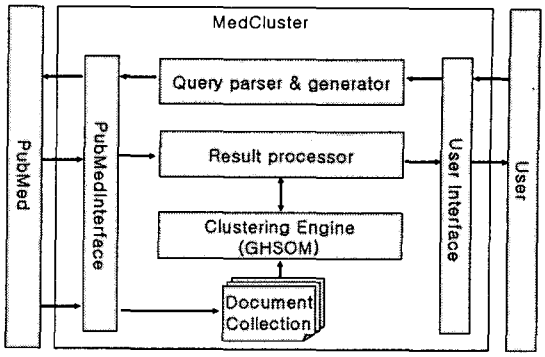
$MQE_m < \tau_m \bullet mqe_0$ 이 되어 한 계층의 성장이 종료되면, 이 맵은 다음 계층으로 확장을 시도한다. 이때 매우 높은 mqe를 가진 이 유닛들은 다음 계층의 새로운 맵에 추가된다. 그리고 각 유닛 i는 (식 6)과 같은 조건을 만족하면 확장하게 된다.

$$mqe_i > \tau_u \bullet mqe_0 \quad \text{(식 6)}$$

4. 시스템 설계

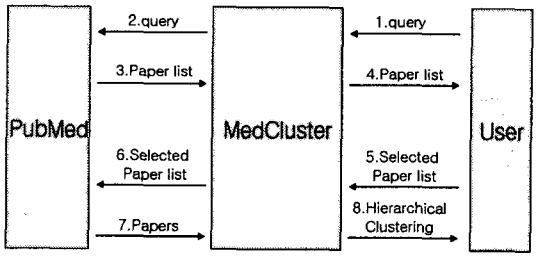
MedCluster 시스템은 기존의 의료문헌정보검색시스템인 PubMed에서 주제어 검색을 하는 이용자들에게 검색 결과 문헌들에 대한 계층적 군집화를 제공함으로써 원하는 문헌 정보에 보다 효율적으로 접근할 수 있도록 설계하였다. MedCluster 시스템의 전체적인 구성은 (그림 2)와 같다.

PubMed Interface는 MedCluster와 PubMed 사이의 데이터를 전달하는 역할을 하고, User interface는 User와 MedCluster사이의 데이터를 전달하는 역할을 한다. Query paser & generator는 User Interface와 pubMed Interface 사이의 질의를 파싱하고 각 모듈에 적합한 형태로 질의를 재생성하여 전달해주는 모듈이다. Result process는 pubMed와 Clustering engine 으로부터 전달된 데이터를 User



(그림 2) 시스템의 구조

Interface로 보내는 역할을 하며, Document Collection은 pubMed로 받아온 문헌들이며, Clustering Engine은 내부적으로 군집화에 필요한 전처리 부분과 실제 군집화를 하는 군집화 모듈로 구성되어있다. 전처리 모듈에서는 각 문헌에 사용된 단어를 기초로 TFIDF를 알고리즘을 이용한 벡터를 파일을 생성한다. 군집화 모듈이 이 시스템의 핵심 모듈로 GHSOM을 사용하여 각 문서간의 유사도를 측정하여 계층적 군집화를 하게 된다.

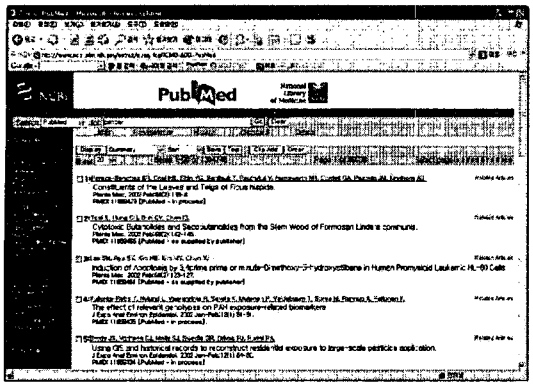


(그림 3) 시스템의 처리과정

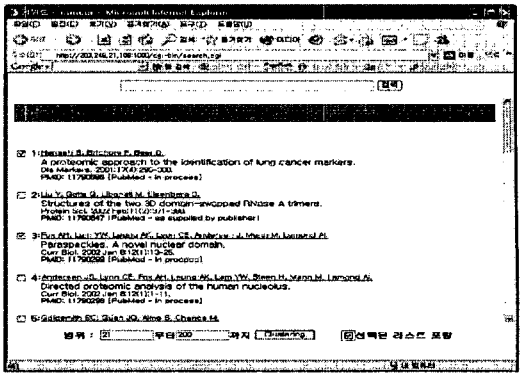
설명한 각 부분간의 흐름을 사용자, 시스템, DB 관점에 보면 (그림 3)와 같이 나타낼 수 있다.

(그림 3)의 각 단계를 설명하면 다음과 같다.

① 사용자의 질의 입력
② 입력된 질의를 파싱하여 데이터 베이스에 전달
③ 데이터 베이스로부터 문헌의 리스트를 받음
④ 문헌 리스트로부터 필요한 부분을 파싱하여 사용자에 보여줌
⑤ 리스트에 관심있는 문헌을 선택하고 범위를 결정
⑥ 선택된 리스트를 데이터베이스에 전달
⑦ 선택된 문헌을 저장
⑧ 저장된 문서를 계층적 군집화를 하여 사용자에 보여준다.



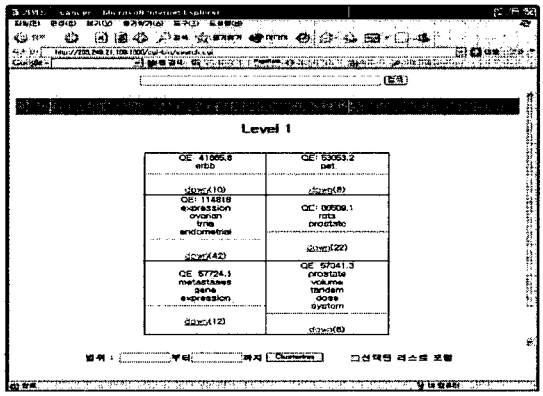
(그림 4) PubMed의 검색화면



(그림 5) MedCluster의 문헌 선택 화면

5.구 현

MedCluster 시스템은 리눅스(Linux) 환경에서 Perl 프로그래밍 언어와 libwww-perl 라이브러리, 그리고 SOMLib 소프트웨어 툴들을 이용하여 구현하였다. (그림 4)는 pubMed의 기본 검색화면이다. 아래의 그림에서도 알 수 있듯이 cancer라는 키워드를 입력하면 무수히 많은 결과를 나타내는 것을 알 수 있다. 이러한 결과를 줄이기 위해서 키워드들 더 사용해서 결과의 폭을 줄일 수 있지만, 사용자에게 있어서 키워드의 선택은 쉽지가 않고 여러번 검색해야 원하는 결과를 찾을 수 있다. (그림 5)은 사용자가 PubMed로부터 검색된 문헌 리스트에 대해 문헌을 하나씩 직접 체크표시를 하거나 아니면 범위를 지정함으로써 군집화를 원하는 문헌들을 선택하는 화면이다. (그림 6)은 선택된 문헌들에 대해 GHSOM을 이용한 계층적 군집화가 이루어진 후 사용자에게 제시되는 결과 화면이다. 이 그림을 통해 검색문헌들이 크게 최상위 계층의 6개 군집으로 나뉜 것을 볼 수 있으며, 각 군집에 포함된 문서의 수는 괄호안에 표시된 것과 같다. 각 군집을 나타내는 셀을 클릭하면 한 단계 아래의 세부 군집들을 화면에 표시해주며, 맨 아래 계층인 말단 군집 셀에서는 대응되는 특정 문헌들에 대한 링크를 제공한다.



(그림 6) MedCluster의 군집화 결과 화면

6. 결론

본 논문에서는 신경망 GHSOM의 구조와 특성에 대해 간략히 살펴보았으며, GHSOM을 채용한 의료문헌 군집화시스템 MedCluster의 설계와 구현에 대해 설명하였다..

참고문헌

[1] D. Merkl and A. Rauber, "The SOMLib Digital Library System". Proc of ECDL'99, 1999

[3] Krista Lagus, "Text Mining with WEBSOM", ACTA Polytechnica Scandinavica, Mathmatics and Computing Series, No.110, 2000

[4] Micheal Dittenbash, Dieter Merkl and Andeas Rauber, "The Growing Hierarchical Self Organizing Map", Proc of IJCNN2000, pp.15-19, 2000

[5] Micheal Dittenbash, Dieter Merkl and Andeas Rauber, "Recent Advance with the Growing Hierarchical Self Organizing Map", Springer - Verlag, 2001