코사인 유사도 기법을 이용한 뉴스 추천 시스템

김상모○, 김형준, 한인규
국민대학교 임베디드 시스템 연구실
simon@koama.net, dndnwjs@gmail.com, in1004kyu@gmail.com

SNS news Recommendation by Using Cosine Similarity

Simon Kim○, Hyung-Jun Kim, In-Kyu Han
Kookmin University, Embedded System Lab.

요 약

사용자별로 SNS/RSS 구독 뉴스 분석을 통해 사용자가 관심이 있는 새로운 뉴스를 추천해 주는 시스템을 설계하고 구현한다. 뉴스 추천 시스템의 설계를 위해 전체 시스템에서 사용자와 서버에서의 작업을 명세하고, 이중에 주요 기능을 담당하는 부분을 구현한다. 구현된 주요 기능은 선호 문서가 들어왔을 때 특징을 추출하고 이를 저장하는 것과 새로운 문서가 들어왔을 때 선호 문서군과 얼마나 유사한지 판별하여 문서에 대한 추천 여부를 결정하는 것이다. 선호 문서의 특징 추출에 대해서는 형태소 분석을 통해 단어와 빈도를 추출하고 이를 누적하여 저장한다. 또한, 새로운 문서가 들어왔을 때 코사인 유사도를 계산하여 사용자가 선호하는 학습문서와의 유사도 비교를 통해 문서 추천 여부를 결정한다. 구현된 시스템에서 실제로 연관된 선호 문서군을 학습시키고, 연관된 새로운 문서 혹은 연관되지 않은 새로운 문서에 대한 추천 여부를 비교하는 것으로 시스템 정확도를 파악한다.

주제어: Naïve Baysian, 코사인 유사도, 형태소 분석, 문서 벡터

1. 서론

오늘날 사용자들은 각종 매체에서 많은 뉴스를 접한다. 이중 원하는 뉴스를 골라보기 위해서 전체 뉴스들을 살펴 보아야 하는 번거로움이 있다. 이 때문에 자동적으로 사용자가 좋아할 만한 뉴스를 제공해 주는 서비스가 구글에서 제공됐다[1]. 하지만 이는 사용자가 어떤 뉴스를 클릭해서 들어갔는지에 대한 패턴 조사에 따른 추천 시스템이다. 단순히 뉴스를 본 것에 그치는 것이 아니라 사용자 자신이 Facebook의 "좋아요"를 클릭한 행동에 따라 추천해 준다면 더 정확한 뉴스의 추천이 가능할 것이다. 또한 현재 Facebook, Google Plus와 같은 소셜 네트워크를 통한 뉴스의 제공이 늘어나고 있는 추세이다. 그림 1에서 보는 것처럼 소셜 네트워크로부터의 뉴스 제공량이 2010에서 2012년까지 10%의 상승폭을 보이고 있다[2]. 이러한 이유로 인해 사용자 선호도 측정에 사용할 뉴스를 소셜 네트워크에서 제공되는 것을 기준으로 잡고 타 소셜 네트워크(Twitter)와 RSS에서 제공되는 뉴스에서 사용자가 좋아할만한 뉴스를 추천해서 보여주는 시스템을 개발하고자 한다.

여기서 해당 시스템의 사용자가 10만 명이라고 가정할 때 각 사용자의 평균 뉴스 구독 수를 5개, 사용자가 업로드하는 뉴스를 20개라고 가정하면, 하루에 생성되는 뉴스의 양은 약 1,000만개가 된다. 생성된 모든 뉴스 중 사용자의 관심을 끌만한 뉴스를 일일이 찾아본다는 것은 매우 번거러운 작업이다. 하지만 이미 선호하는 뉴스에 대한 정보는 SNS/RSS 구독 정보를 통해 얻을 수 있다. 따라서 우리는 각 사용자 별 SNS/RSS 구독 정보에 대한 분석을 통해 많은 양의 새로운 뉴스가 생성되었을 때 이 중 사용자가 관심을 가질 가능성이 높은 뉴스를 기존에 관심이 있었던 문서들과의 유사도를 기준으로 추천해 주는 시스템을 설계하고 구현하고자 한다.

More Americans Getting News Digitally and from Social Networks

Where did you get news yesterday?	2010 %	2012 %	Change
NET Online/Mobile*	34	39	+5
Online	34	34	0
Cell, tablet, other mobile	--	17	--
Social networking sites like Facebook, Google Plus	9	19	+10
Twitter	2	3	+1
Email	14	16	+2

그림 1. 2012 News Consumption Survey

2. 관련 연구

본 연구에서의 주요 기능들은 자연어처리 기법을 기반으로 구현되었다. 문서의 내용을 분석하고자 대상 어절의 모든 가능한 분석 결과를 출력하는 형태소 분석이나 문서 내용 분석을 기반으로 문서의 범주를 이미 정의된 범주에 할당하는 문서 분류 같은 것이 이에 해당된다. 형태소란 자연언어에서 '의미를 갖는 최소의 단위'이며 형태소 분석은 하나의 말마디를 여러 형태소로 분리한 후에 형태적 변형을 처리하는 것이다[3,4,5].

문서 분류는 인터넷 상의 블로그나 SNS 등 때문에 증가하는 문서의 양으로 인해. 효율적이며 정확한 문서 범주화를 요구하게 되며 발전하였다[6,7,8,9]. 문서 내용에 등장하는 용어에 기초한 통계적 방법, 문장의 의미를 파악하는 의미적 해석 방법, 문장 구조에 의존하는 구조적 방법으로 구별된다. 문헌에 나타난 용어의 특성 분석을 바탕으로 하는 자동 분류 시스템의 연구가 가장 활발하게 이루어지고 있다. 또한, 한글과 영어는 형태론적 구조가 다르므로 형태소 추출 과정은 다르지만 추출된 용어의 빈도 등의 특성 분석은 본질적으로 유사하다[10,11].

문서 분류는 문서 내용 분석을 통해 용어 집합을 통해 문서를 분류하는 방법과 분류 모델을 생성하여 문서를 분류하는 방법이 있으며, 또한 문장구조를 이용하는 방법도 존재한다[12,13]. 본 연구에서는 뉴스가 하나의 문서이기 때문에 사용자가 선호할 뉴스와 선호하지 않을 뉴스에 대한 분류 기술이 필요하다. 문서 분류 기법 중에서 나이브 베이지안 방법은 조건부 확률을 단순 확률 곱 연산과 나눗셈 연산으로 간단화시키는 방법이다. 어떠한 문서가 있을 때 그 문서가 추천하는 클래스에 속해 있을 확률이 조건부 확률이다. 이 확률이 높을수록 해당 문서는 추천할 만하다고 판단한다. 이 확률을 쉽게 구하기 위해 Naïve Bayesian을 이용하면 독립적인 곱셈 연산과 나눗셈 연산으로 단순화시키는 벡터 연산이 가능하다.

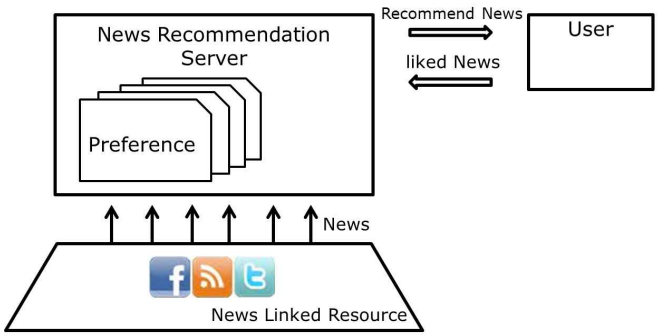

그림 2. 뉴스 추천 시스템의 서비스 구조

3. 뉴스 추천 시스템 구조

본 연구에서 설계한 시스템의 서비스 구조는 그림 2와 같다. 사용자는 단말에서 서버가 제공하는 2가지 서비스를 받는다. 추천된 뉴스를 확인하거나, 자신이 선호하는 뉴스를 선택할 수 있는 것이다. 그리고 서버는 개인별로 선호하는 뉴스에 대한 데이터베이스를 가지고, 새로운 뉴스를 SNS/RSS를 통해 얻고, 이 중 사용자가 관심을 가질 만한 뉴스를 추천해 주는 것이다.

중요 구성요소 별 시나리오는 크게 2가지로 나뉜다. 먼저 사용자가 선호 뉴스를 등록하면 선호 뉴스로 등록된 뉴스는 문서로 간주하고 해당 문서에 대한 특징 벡터를 추출한다. 추출된 특징 벡터는 개인 데이터베이스에 누적되어 저장된다. 매일 새로운 뉴스가 업로드되면 개인별 데이터베이스의 누적된 문서 특징 벡터를 이용하여 추천 여부를 판별한다. 새로운 뉴스의 내용 또한 특징 벡터를 추출하고 누적된 문서 특징 벡터와 유사도를 비교하여 추천 여부를 판단한다. 추천 여부에 따라 개인별로 관심을 가질 만한 문서가 추천된다.

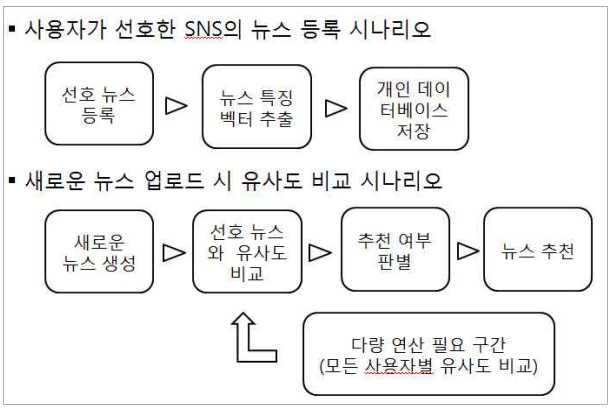

그림 3. 뉴스 추천 서비스 시나리오

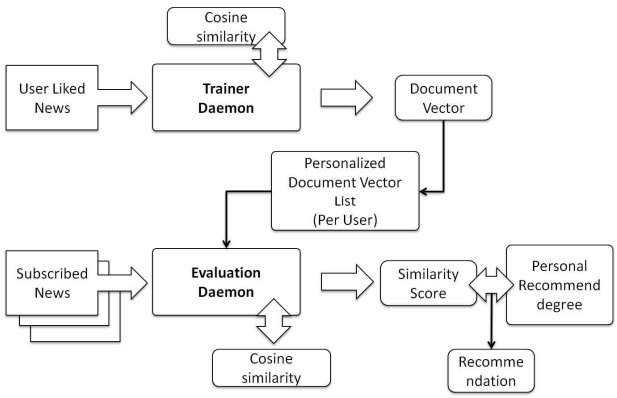

그림 4. 뉴스 추천 시스템의 구조

3.1 뉴스 추천 시스템의 구조

그림 4는 중요 구성 요소를 구현하기 위한 전체 시스템 구조를 나타낸다. 사용자가 선호한다고 표시한 뉴스를 User Liked News로 정의하고, RSS/SNS로부터 유입되는 뉴스를 Subscribed News라 정의한다. User Liked News를 입력 문서로 하는 Trainer Daemon 컴포넌트는 특징 벡터(한 문서에서 사용된 단어와 해당 단어의 개수의 매핑 벡터)인 Document Vector를 추출하여 Personalized Document Vector List(PDVL) 라는 데이터 저장소에 저장한다. 이 저장소는 각 개인 별로 존재하며 이를 기반으로 사용자 별 추천여부를 계산한다. 매일 새로 생성되는 Subscribed News와 PDVL을 기반으로 Evaluation Daemon 컴포넌트에서는 새로운 뉴스에 점수를 매긴다. 계산된 점수는 선호 뉴스들과의 유사도를 나타내는 Similarity Score로 정의된다. 이 점수와 사용자 별 유사도 정도를 나타내는 Personal Recommend degree와 비교하여 추천 여부를 결정한다.

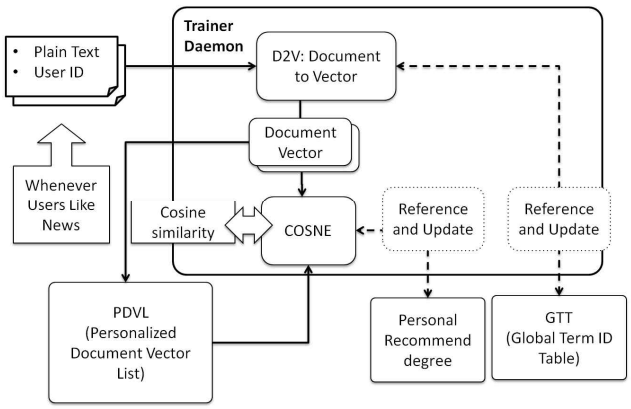

그림 5. 학습 데몬의 내부구조도

3.2 학습 데몬

학습 데몬(Trainer Daemon)은 사용자가 선호하는 문서에 대한 정보들을 수집하고 저장하는 모듈이다. 사용자가 선호하는 뉴스가 들어오면 이는 문서로 취급되며 D2V(Document to Vector)라 불리는 문서에서 특징 벡터를 추출하는 모듈에 입력된다. D2V 모듈은 문서 내부 단어들을 고유 ID값과 단어의 빈도수를 나타내는 COUNT값으로 이루어진 벡터인 Document Vector로 만들어서 PDVL에 저장한다. 모듈은 필수 기능 외에 고유 ID값을 다루는 모듈이 필요하다. 따라서 시스템 전체 단어를 관리하는 테이블이 필요하다. 이 테이블을 다루는 기능을 담당하는 모듈은 Global Term ID Table(GTT)로 정의하였다.

D2V는 문서를 벡터화하는 과정에서 GTT를 참조하며, 새로운 단어가 들어오면 GTT를 업데이트 하여 새로운 ID값을 제공받는다. Training을 하고 나서 현재 PDVL의 데이터와 선호 뉴스와의 유사도 계산 방식으로 뉴스 추천 기준값 Similarity Score를 계산한다. 이 기준값은 이후 새로 들어온 뉴스와의 유사도 계산 점수와 비교하는 기준값이 된다. 해당 점수는 PDVL과 사용자가 선호하는 뉴스와의 유사도 점수이기 때문에 이보다 높으면 추천 정도가 높다고 볼 수 있다. 뉴스 추천 기준값 Similarity Score는 개인별 데이터베이스에 저장한다. 만약 저장된 Score가 있다면 둘 사이의 평균값을 계산하여 새로 저장한다. 이 점수를 개인별 추천 정도를 나타내는 Personal Recommend degree라고 정의한다.

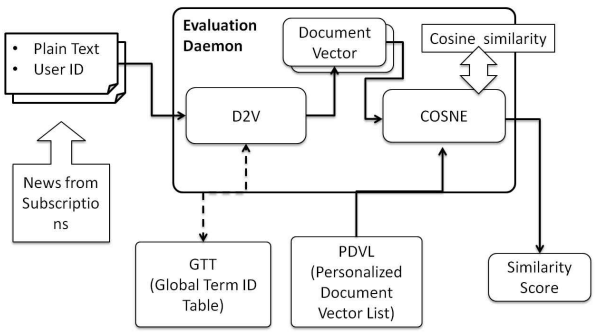

그림 6. 평가 데몬의 내부 구조도

3.3 뉴스 추천 및 평가

Evaluation Daemon은 새로 생성된 뉴스가 사용자가 관심을 가질 만한 뉴스인지 판별하는 모듈이다. 새로운 뉴스가 업데이트 되는 것을 감지하여, 문서를 가져오는 기능은 구현하지 않는다. 새로운 뉴스가 업데이트 되었다고 가정하고 다양한 문서들을 모듈의 입력 데이터로 사용한다. 입력된 데이터는 D2V를 이용하여 특징 벡터를 추출하고, PDVL과 함께 COSNE 모듈에서 코사인 유사도 연산을 통해 선호 뉴스들과의 유사도 정도를 도출해 낸다. 이 값은 Similarity Score로 정의되며 저장된 Personal Recommend degree와 비교하여 추천여부를 결정한다. Personal Recommend degree 보다 높으면 강력한 추천으로 추천하며, Recommend degree의 50% 보다 높은 상위 10개의 문서는 보통 추천으로 추천한다.

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta$$

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

그림 7. 코사인 유사도 계산식

PDVL과 새로 들어온 뉴스의 Document Vector를 가지고 문서 분류 연산 중 하나인 코사인 유사도 알고리즘을 사용하여 추천 문서를 추천하는 모듈이다. 코사인 유사도는 그림 7과 같은 공식을사용한다. 두 백터 a, b 사이의 각도를 나타내는 $cos(\Theta)$는 두 벡터 a와 b 사이의 각도를 나타내기 때문에 이를 유사도 정도로 볼 수 있다. 여기서 $cos(\Theta)$ 그림-7 의 연산으로 구할 수 있다. 각 문서의 특징을 추출해 낸 D2V는 벡터로 표현이 가능하기 때문에 이와 같은 코사인 유사도 연산에 사용이 가능하다. 본 프로젝트에서 a는 비교할 문서의 벡터, b는 PDVL의 벡터가 되고 $cos(\Theta)$ 는a와 b의 유사도 정도가 된다.

4. 실험 및 성능 평가

성능 측정을 위한 실험 환경은 표 1과 같다.

표 1. 실험 환경

CPU	Intel® Core™ i5-3570 CPU @ 3.40GHz
RAM	8GB DDR3 SDRAM
OS	Ubuntu 12.04

실험 데이터는 주로 IT 관련 뉴스 30개로 Training 시키고 임의의 800개의 뉴스에 대하여 추천 정도를 테스트한다. 실험 결과 학습에 사용한 30개의 뉴스와 추천 기준값이 되는 유사도보다 높은 6개의 뉴스를 선호도가 매우 높다고 판단되는 강력한 추천 대상으로 하였다. 추천 기준값의 50%보다 높은 뉴스는 약 300개 정도가 나왔고 그 중 상위 10개를 보통 추천 뉴스로 하였다. 그리고 나

머지 뉴스는 추천 대상에서 제외하였다. 추천 대상이 된 뉴스의 주제는 주로 삼성, LG, 구글과 같은 IT 업계와 관련된 뉴스였으며 추천 기준값의 50% 보다 높은 뉴스 중 상위 10개의 뉴스들 중 6개는 IT와 관련된 주제의 뉴스였지만, 나머지 4개는 IT 용어가 나오긴 하지만 경제나 사회에 관한 뉴스였다. 그 외 기준 유사도 점수의 50% 이상에 해당 하는 뉴스 중 상위 10개의 뉴스를 제외한 뉴스 들은 IT와 관련된 뉴스가 있긴 했지만 대부분 정치, 경제, 사회에 대한 뉴스였다. 결과적으로 유사도 점수가 높게 나온 뉴스는 대부분 사용자가 선호할 만한 뉴스였고, 낮게 나온 뉴스는 대부분 선호할 만한 뉴스가 아니였다.

표 2. 뉴스 추천 시스템의 실험 결과

추천 정도	강력 추천 (기준 유사도 이상)	보통 추천 (기준 유사도의 50% 이상 중 상위 10개)		유사도의 50 % 이상	비추천 (그 외)
추천된 뉴스	6개 (IT 회사, 제품)	6개 (IT 회사, 제품)	4개 (경제, 사회 - IT 관련정책)	300개 (IT, 기타)	나머지 기타

5. 결론

본 연구에서는 사용자가 좋아할 만한 뉴스를 가져오면 유사도 점수를 계산하여 새로 유입되는 뉴스와의 유사도 점수와 비교하여 추천하는 시스템을 구현하였다. 단순히 문서의 단어에 대한 빈도 계산만 가지고도 어느 정도 추천하는 시스템이 완성되었기 때문에 단어 빈도뿐만이 아닌 단어의 가중치를 고려한다면 더욱 정교한 추천 시스템이 구현될 것이다. 여기에 뉴스를 수집하는 모듈을 추가하여 실시간으로 추천되는 시스템으로 확장하면, 실제 서비스가 가능해지고 이를 웹페이지, 안드로이드, 아이폰 앱으로 확장이 가능할 것이다.

참고문헌

[1] Personalized News Recommendation Based On Click Behavior, http://googlenewsblog.blogspot.kr/2011/04/automatic-personalization-and.html

[2] Trend in News Consumption 1991-2012, http://www.people-press.org/

[3] 강승식, "다층 형태론과 한국어 형태소 분석 모델", 제6회 한글 및 한국어 정보처리 학술발표 논문집, pp.140-145, 1994.

[4] 강승식, 이하규, 손소현, 홍기채, 문병주, “조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법”, 한국정보과학회 가을 학술발표논문집, 28권 2호, pp. 196-198, 2001.

[5] 최재혁, "형태소 분석을 통한 한영 자동 색인어 추출", 정보과학회논문지(B), 제23권, 제12호, pp. 1279-1288, 1996.

[6] Yang, Y. and Xin Liu, “A Re-examination of Text Categorization Methods”, In Proc. of Conference on Research and Development in Information Retrieval(SIGIR 99), pp.42-49, 1999.

[7] Y. Yang and J. P. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” In Jr. D H. Fisher(Ed.), 14th International Conference on Machine Learning, Morgan Kaufmann, pp.412-420, 1997.

[8] F. Sebastiani, “Machine Learning in Automated Text Categorization”, Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 1999.

[9] V. Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag, 1995.

[10] 고영중, 박진우, 서정연, “문장 중요도를 이용한 자동 문서 범주화”, 정보과학회 논문지: 소프트웨어 및 응용, 29권 6호, pp.417-423, 2002.

[11] 이지행, 조성배, “전자우편 문서의 자동분류를 위한 다중 분류기 결합”, 정보과학회 논문지: 소프트웨어 및 응용, 29권 3호, pp.192-201, 2002.

[12] 이경찬, 강승식, “용어 가중치와 역범주 빈도에 의한 자동 문서 범주화”, 제15회 한글 및 한국어 정보처리 학술발표 논문집, pp.14-17, 2003.

[13] K. C. Lee, S. S. Kang, K. S. Hahn, “A Term Weighting Approach for Text Categorization”, AIRS 2005, pp.673-678, 2005.