위키피디아를 이용한 반자동 학습 기반의 cQA 서비스 주제 분류 시스템

김태현○
동아대학교 컴퓨터 공학과
karvien@gmail.com

A Topic Classification System in cQA Services Based on Semi-Automatic Learning Using Wikipedia

Taehyun Kim○
Computer Engineering, Dong-A University

요 약

본 논문은 커뮤니티 기반의 질의-응답 서비스에서 사용자 질의의 주제를 분류하는 시스템을 소개한다. 커뮤니티 기반의 질의-응답 서비스는 분야에 따라 다양한 주제를 가질 수 있으며 오늘 날 사용자 질의의 주제 분류에는 통계 기반의 분류 방법이 많이 이용되고 있다. 통계 기반의 분류 방법으로 사용자 질의를 분류하기 위해서는 주제에 적합한 대량의 학습 말뭉치가 필요하다. 주제에 적합한 대량의 학습 말뭉치를 사람이 직접 구축하는 것은 많은 시간과 비용이 든다. 따라서 본 논문에서는 이러한 문제를 해결하기 위해 위키피디아 문서를 Supervised K-means Clustering 기법으로 주제별로 분류함으로써 학습 말뭉치를 반자동으로 구축하는 방법을 제안한다. 그 다음, 생성된 학습 말뭉치로 지지 벡터 기계를 학습하여 사용자 질의의 주제를 분류하게 된다. 위키피디아 문서와 사용자 질의는 다른 도메인의 문서임에도 불구하고 본 논문의 시스템으로 사용자 질의의 주제를 분류한 결과 77.33%의 정확도를 보였다.

주제어: cQA, 주제 분류, 위키피디아, 반자동 학습 말뭉치 구축

1. 서론

오늘 날 인터넷이 발전하면서 NAVER 지식IN, Yahoo! Answer, Live QnA와 같이 사용자가 자신이 원하는 정보에 대한 질의를 하면 다른 사용자가 이에 응답하는 커뮤니티 기반의 질의-응답 서비스(community-based Question Answering Service, 이하 cQA)의 중요성이 늘어나고 있다.[1] 본 논문에서는 이러한 cQA 서비스에서 사용자 질의의 주제(Topic)를 분류하는 시스템을 소개한다.

질의의 주제를 분류하는 것은 응답자가 자신의 전문 분야를 선택하여 효과적으로 응답을 할 수 있게 도와줄 뿐만 아니라, 사용자가 질의를 등록할 때에도 작성한 질의에 대해 주제를 추천함으로써 사용자에게 주제 선택의 편의성과 정확성을 제공한다.

cQA 서비스와 유사한 QA 서비스에서 질의의 주제를 분류하기 위해서는 규칙 기반의 분류 방법과 통계 기반의 분류 방법 등이 있으며, 일반적으로 통계 기반의 분류 방법을 많이 이용한다.[2] cQA 서비스는 목적에 따라 다양한 주제를 가지며 통계 기반의 방법으로 질의의 주제를 분류하기 위해서는 해당 cQA 서비스의 주제에 적합한 대량의 학습 말뭉치가 필요하다. 하지만 주제에 적합한 대량의 학습 말뭉치를 사람이 직접 구축하는 것은 많은 시간과 비용이 든다. 따라서 본 논문에서는 많은 시간과 비용의 필요 없이 위키피디아를 이용하여 일부의 분류된 문서만으로 대량의 학습 말뭉치를 반자동으로 구축하고 이를 이용하여 사용자 질의의 주제를 분류하는 시스템을 제안한다.

본 논문에서 학습 말뭉치 구축에 위키피디아를 이용한 이유는 다음과 같다. 첫째, 위키피디아는 온라인 백과사전으로 지리, 역사, 종교, 사회, 스포츠 등 생활 전 분야에 대해 다양하게 다루고 있어 구축하려는 cQA 서비스의 주제에 적합한 학습 말뭉치를 구축 할 수 있다는 이점이 있다. 둘째, 위키피디아 문서는 시간이 지남에 따라 증가하고 있으며 이를 이용하면 많은 시간과 비용의 투자 없이 학습 말뭉치를 확장하고 업데이트 할 수 있다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 질의 주제 분류 분야에 수행되었던 관련 연구들에 대해 살펴보고, 3장에서는 사용자 질의 주제 분류 시스템에 대해서 설명한다. 그리고 4장에서는 실험 및 평가를 하며, 5장에서는 결론 및 향후 과제를 기술한다.

2. 관련 연구

2-1. 질의 주제 분류에 관한 연구

질의 주제 분류에 관한 연구는 [2-3]과 같은 QA 시스템의 연구가 주를 이루어왔으며 최근에는 [1], [8-10]과 같은 cQA 시스템의 질의 주제 분류 연구도 있었다. QA 시스템과 cQA 시스템의 질의 형태적 차이점은 QA 시스템의 질의는 형태가 제한적이고, cQA 시스템의 질의는 형태가 제한적이지 않다는 것이다.[1] 하지만 두 시스템 모두 질의 분류라는 공통점이 있고 질의 분류는 문서 분류(Document Classification)의 한 형태이므로 QA 시스템과 cQA 시스템의 질의 분류 연구는 관련이 깊다.[5-7]

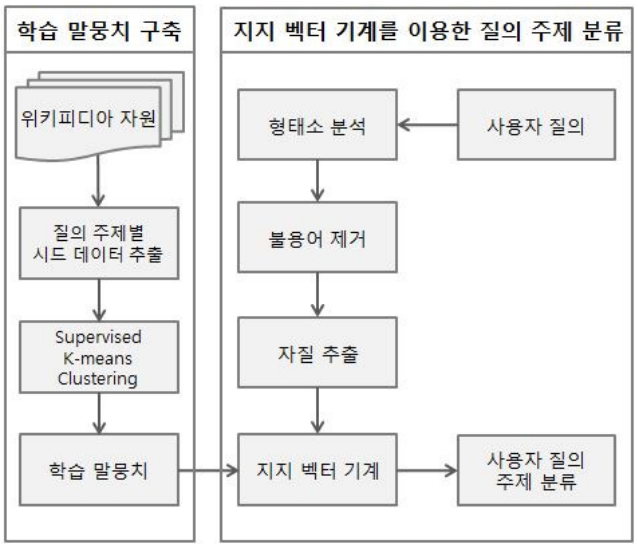
2-2. 위키피디아를 이용한 분류 연구.

위키피디아를 이용한 분류 연구로는 [4,10] 등 이 있었다. [4]는 위키피디아를 이용해 트위터 문서의 주제를 분류하는 연구로써, 트위터는 문서의 형태가 제한적이지 않다는 점에서 cQA 시스템의 질의와 유사하다고 볼 수 있다. [10]은 cQA 시스템에서 질의의 주제를 분류하는 연구로써, 위키피디아의 구조적 특징인 링크 정보를 이

용함으로써 질의의 주제를 분류하였다.

3. 사용자 질의 주제 분류 시스템

이 절에서는 위키피디아를 이용하여 학습 말뭉치를 구축하고 지지 벡터 기계를 학습하여 사용자 질의의 주제를 분류하는 방법에 대해서 설명한다. 전체적인 시스템의 구성도는 아래의 <그림 1>과 같다.



<그림 1> 전체적인 시스템 구성도

3.1. 학습 말뭉치의 반자동 구축

본 논문의 시스템에서는 학습 말뭉치를 구축하기 위한 지식 베이스로 위키피디아 자원을 이용하였다. 학습 말뭉치는 위키피디아 문서를 Supervised K-means Clustering 기법을 이용하여 분류함으로써 반자동으로 구축하였다.

3.1.1. 질의 주제별 시드 데이터 추출

시드 데이터(Seed Data)는 K-means 알고리즘의 초기 중심점을 선정할 때 각 중심점이 해당 주제를 대표하도록 함으로써 문서를 원하는 주제별로 분류 할 수 있도록 한다. 시드 데이터는 사람이 직접 판단하여 해당 주제를 대표하는 일부 문서를 선정한다. 이를 통해 일부의 문서만으로 위키피디아 전체 문서를 cQA 시스템에서 원하는 주제별로 분류하여 학습 말뭉치를 반자동으로 구축할 수 있다.

3.1.2. Supervised K-means Clustering 기법을 이용한 위키피디아 문서 분류

기존의 K-means 알고리즘의 경우 중심점 초기화 방법은 Random Partition 방법, Forgy Algorithm, MacQueen 알고리즘, Kaufman 알고리즘 등이 있다. 이와 같은 중심점 초기화 방법은 데이터의 클러스터와 무관하게 중심점이 무작위로 선정된다. 본 연구에서는 cQA 시스템에서 선정한 주제별로 시드 데이터를 부여하고 이를 이용하여 지도함으로써 주제별로 적절한 초기 중심점을 할당하게 된다. 그 후 기본적인 K-means 알고리즘을 적용하여 위키피디아 전체 문서를 시스템에서 선정한 주제별로 분류한다.

3.1.3. 분류 된 위키피디아 문서를 이용한 학습 말뭉치 구축

이전 절에서 시드 데이터와 Supervised K-means Cluster-ing 기법을 이용하여 위키피디아 문서를 분류하였다. 이렇게 분류된 문서는 각 주제별로 분류되었지만 양질의 학습 말뭉치를 구축하기 위해서 각 주제별 중심점과 가장 코사인 유사도(Cosine Similarity)가 높은 적정량의 문서들만을 선정한다. 그 후 선정한 문서들을 이용하여 학습 자질을 추출하고 최종적으로 학습 말뭉치를 구축하였다.

3.2. 지지 벡터 기계를 이용한 사용자의 질의 주제 분류

지지 벡터 기계는 통계 기반의 분류 방법으로 안정적이고 높은 성능을 제공하는 분류기이지만 대량의 학습 말뭉치를 필요로 한다는 문제점이 있다. 본 논문에서는 위키피디아 전체 문서를 최소한의 시드 데이터만을 이용하여 분류하였고, 이를 이용하여 학습 말뭉치를 반자동으로 구축함으로써 이러한 문제점을 해결하였다.

지지 벡터 기계를 이용하여 사용자 질의를 분류하기 위한 단계는 다음과 같다. 첫째, 사용자 질의를 입력받고 형태소 분석을 통해 질의를 형태소 단위로 분리하였다. 둘째, 형태소로 분리된 질의에서 “해결”, “문제”, “질문”과 같이 불용어에 가까운 단어는 불용어 목록을 이용하여 제거하였다. 셋째, 분류에 필요한 자질만을 추출하였다. 마지막으로, 추출한 자질과 지지 벡터 기계를 이용하여 사용자 질의의 주제를 분류하였다.

4. 실험 및 평가

이 절에서는 “Computer, Entertainment, Health, Society, Sport, Study"와 같은 6가지 질의 주제에 대해서 실험하였다. 또한 다음과 같은 두 가지 항목을 기준으로 본 논문의 질의 분류 시스템을 실험하고 평가 하였다.

- **Supervised K-means Clustering 기법을 이용한 위키피디아 문서의 주제별 분류 평가**
- **지지 벡터 기계를 이용한 질의의 주제 분류 평가**

4.1. Supervised K-means Clustering 기법을 이용한 위키피디아 문서의 주제별 분류 실험

지지 벡터 기계를 이용하여 질의의 주제를 정확하게 분류하기 위해서는 양질의 학습 말뭉치가 필요하다. 따라서 양질의 학습 말뭉치를 구축하기 위해 Supervised K-means Clustering 기법을 이용하여 위키피디아 문서들을 주제별로 분류하였고, 그 중 각 주제별로 중심점과 코사인 유사도가 높은 일부 문서들만을 선정하였다. 이 실험은 Supervised K-means Clustering 기법으로 분류한 문서들이 정확하게 분류되었는지 평가하는 실험이다.

4.1.1. 실험 데이터

본 실험에서 사용한 위키피디아 덤프에 포함된 문서의 개수는 약 78만개 정도였으며 이 중 “위키”, “틀”, “숫자”, “년”, “월”과 같은 문서를 제외하고 약 19만개의 문서를 이용하였다.

Supervised K-means Clustering 기법에서 초기 중심점을 할당하기 위한 시드 데이터는 각 주제별로 사람이 직접 선정하였으며, 해당 주제별로 800KB 정도의 용량을 기준으로 하였다. Supervised K-means Clustering 기법의 자질로는 명사만을 추출하였으며 명사의 TFIDF 가중치를 이용하였다. 그리고 양질의 학습 말뭉치를 구축하기 위해 각 주제별로 중심점과 가장 코사인 유사도가 높은 3100개의 문서를 추출하였다.

본 실험에서는 테스트 데이터로 각 주제에 대해 추출한 3100개의 문서들 중 100개의 문서를 TFIDF 가중치 별로 고르게 추출하여 해당 문서가 정확하게 분류되었는지

에 대한 정확도(Accuracy)를 측정하였다.

4.1.2. 실험결과 및 평가

<표 1> Supervised K-means Clustering 기법을 이용한 문서 분류 성능

주제(Topic)	정확도(Accuracy)
Computer	72%
Entertainment	77%
Health	80%
Society	85%
Sport	90%
Study	68%
평균	78.66%

위의 <표 1>은 각 주제별로 위키피디아 문서를 분류한 결과이다. 분류 자질로써 명사의 TFIDF 가중치만을 사용하였음을 고려할 때 추가적인 자질을 적용한다면 더 높은 성능을 기대할 수 있다.

4.2. 지지 벡터 기계를 이용한 질의의 주제 분류 실험

본 실험은 각 주제별로 테스트 질의를 선정하고 지지 벡터 기계를 이용하여 질의의 주제를 분류하였을 때 본 논문의 시스템이 얼마나 정확하게 분류하는지에 대한 실험이다.

지지 벡터 기계의 학습과 테스트에서 이용되는 분류자질로써는 실험 4.1. 과 동일하게 명사의 TFIDF 가중치를 사용하였다.

4.2.1. 실험 데이터

학습 데이터로는 본 논문에서 제안한 방법으로 구축한 학습 말뭉치를 이용하였다. 학습 말뭉치는 총 18600개의 문서로 이루어져 있으며 각 질의 주제 별로 3100개의 문서를 포함한다.

테스트 데이터로는 본 논문의 cQA 시스템에서 사용하는 6개의 질의 주제에 대해서 NAVER 지식IN에서 실제 사용자들의 질의를 추출하였다. 각 질의 주제 별로 50개의 질의를 추출하였으며, 이렇게 추출한 총 300개의 질의를 지지 벡터 기계를 이용하여 주제 분류를 하였다.

4.2.2. 실험결과 및 평가

<표 2> 지지 벡터 기계를 이용한 질의 주제 분류 성능

주제(Topic)	정확도(Accuracy)
Computer	92%
Entertainment	70%
Health	80%
Society	76%
Sport	78%
Study	68%
Average	77.33%

위의 <표 2>는 각 주제에 해당하는 사용자 질의의 주제를 본 논문의 시스템으로 분류 한 결과이다. [1]의 연구에서 제안한 질의 분류 시스템에 비해서 본 논문의 분류 시스템이 더 높은 성능을 제공하는 것은 아니지만 질의 주제에 적합한 학습 말뭉치를 반자동으로 구축했다는 점과 분류에 기본적인 자질만 사용했다는 점을 고려할 때 유의미한 실험이라고 볼 수 있다.

5. 결론 및 향후 과제

본 논문은 지속적으로 증가하고 다양한 분야의 정보를 다루고 있는 위키피디아를 이용하여, 서비스하려는 cQA 시스템의 주제에 맞는 학습 말뭉치를 반자동으로 구축하고 이를 이용하여 지지 벡터 기계로 사용자 질의의 주제를 분류하는 방법을 제안하였다. 제안한 방법으로 사용자 질의의 주제를 분류하였을 때 77.33%의 정확도를 얻을 수 있었다. 본 논문에서 제안하는 주제 분류 시스템이 관련 연구의 다른 시스템보다 높은 성능을 제공하는 것은 아니지만 질의 주제에 맞는 학습 말뭉치를 반자동으로 구축했다는 점과 기본적인 분류 자질만 사용했다는 점을 고려하면 유의미한 실험이라고 볼 수 있다. 또한 분류에 다양한 자질과 기법들을 적용한다면 더 높은 성능을 기대할 수 있을 것이다.

향후 과제로는 본 논문의 시스템에서 사용한 기본적인 명사 자질 뿐만 아니라, 추가적인 자질들을 시스템에 적용하여 더 높은 성능을 내도록 하는 것이 되겠다. 동의어나 약어 사전과 더불어, 위키피디아의 메타 정보인 RedirectPage나 Link 정보를 이용하여 주제 분류 시스템의 성능 향상에 관한 연구를 수행할 것이다.

참고문헌

[1] 배경만, 고영중, and 김종훈. "커뮤니티 기반의 질의 응답서비스 (cQA) 에서 질문-응답 쌍의 구조적 특징을 이용한 언어 모델 기반의 주제 분류 기법." 정보과학회논문지: 소프트웨어 및 응용 39.8 664-671. (2012)

[2] 김학수, 안영훈, and 서정연. "한국어 질의응답시스템을 위한 지지벡터기계 기반의 질의유형분류기." 정보과학회논문지: 소프트웨어 및 응용 30.5·6 466-475. (2003)

[3] 엄재홍, and 장병탁. "대규모 문서 데이터 집합에서 Q&A 를 위한 질의문 분류 기법." 한국정보과학회 2000 년도 봄 학술발표논문집 제 27 권 제 1 호 (B) 27.1B : 253-255. (2000)

[4] 장재영. "한글 위키피디아를 이용한 트위터 문서의 주제별 클러스터링 기법." 한국인터넷방송통신학회 논문지 14.5 : 189-196. (2014)

[5] R. Prasad, P. Natarajan, K. Subramanian, S.Saleem, R. Schwartz, "Finding Structure in NoisyText: Topic Classification and Unsupervised Clustering," Proc. AND. (2007)

[6] D. Liu, S. McVeety, R. Prasad, P. Natarajan, "Semi-supervised Topic Classification for Low Resource Languages," IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp.5093-5096. (2008)

[7] D. B. Bracewell, J. Yan, F Ren, S Kuroiwa, "Category Classification and Topic Discovery of Japanese and English News Articles," Electronic Notes in Theoretical Computer Science, vol.225, no.2, pp.51-65. (2009)

[8] 권순재, et al. "LSP 분류 기법을 이용한 한국어 및 한국어 문법 cQA 시스템." 2014 한국정보과학회 제 41 회 정기총회 및 동계학술발표회 : 1263-1265. (2014)

[9] 연종흠, 심준호, and 이상구. "확장된 나이브 베이즈 분류기를 활용한 질문-답변 커뮤니티의 질문 분류." 정보과학회논문지: 컴퓨팅의 실제 및 레터 16.1 : 95-99. (2010)

[10] CAI, Li, et al. Large-scale question classification in cQA by leveraging Wikipedia semantic knowledge. ACM: In Proceedings of the 20th ACM international conference on Information and knowledge management. p. 1321-1330. (2011)