LDA를 이용한 대용량 블로그 문서 처리

조요한O, 김동우, 문일철, 오혜연
한국과학기술원, 전자전산학과
{yohan.jo, dw.kim}@kaist.ac.kr, icmoon@smslab.kaist.ac.kr, alice.oh@cs.kaist.ac.kr

Massive Blog Data Processing Using LDA

Yohan JoO, Dongwoo Kim, Il-Chul Moon, Alice Haeyun Oh
Dept. of EECS, KAIST

요 약

본 논문은 기계학습 분야에서 주제 분석을 위해 사용되는 문서 모델인 LDA를 수만 개의 블로그 문서에 적용해보고 나타나는 다양한 현상들에 대해 분석을 시도하였다. 우선 인터넷 상의 블로그 데이터를 네 가지 유형으로 분류하여 주제를 분석 하였다. 각 유형의 데이터가 가지는 특징과 그로 인해 나타나는 현상들을 살펴본 뒤에 혼잡도의 개념을 도입하여 학습된 모델들의 성능을 비교하였다. 주제 분석을 통해 블로그 데이터에서 은행, 문학, 인종 등의 주요 주제들이 추출되는 결과를 확인할 수 있었으며, 데이터 유형별 혼잡도 분석을 통해 블로그 데이터를 주제 분류에 적용할 때 고려할 만한 사항들을 논의하였다.

1. 서 론

Technotari의 보고서[1]에 따르면 하루에 약 12만 개의 블로그가 새로 생성되며, 140만 개 이상의 새로운 블로그 글들이 작성되고 있다. 블로고스피어(Blogosphere)의 폭발적 증가 추세는 새로운 사회·문화적 현상을 야기하고 있으며, 어떤 의미에서는 이러한 블로고스피어가 실세계를 반영한다 말할 수 있다. 이러한 측면에서 볼 때, 블로그 상에서 이야기되고 있는 주제들의 경향을 파악하는 것은 실제로 사람들이 어떠한 주제에 얼마만큼의 관심을 가지고 있는지 파악할 수 있는 중요한 척도가 될 수 있다.

경향을 파악하기 위해 선행되어야 하는 일은 문서들의 주제를 분류하는 것이다. 주제 분류 연구는 학술 문서 같은 데이터를 이용해 활발하게 진행되어 왔지만 블로그 문서들은 이런 문서와 큰 차이가 있다. 첫째, 블로그 문서들은 일반적인 문서에 비해 그 내용이 문맥에 의존하는 정도가 심하다. 즉 문서 내에 존재하는 단어의 의미를 해석하기 위해 해당 문서의 내용뿐만 아니라 문서와 연관된 주변지식을 요구하게 되는 경우가 발생하는 것이다. 둘째, 기존 문서보다 훨씬 광범위하고 예측하기 어려운 주제들을 포함하고 있다. 셋째, 블로그 문서들은 정제된 언어와 형태를 가지고 있지 않다. 수천만 명의 사용자들이 서로 다른 주제에 대하여 이야기하는 까닭에 주제를 분석하는 것이 까다로우며, 기존의 연구와 다른 방향을 가질 수밖에 없다.

최근에 블로그 데이터를 이용한 연구가 활발히 진행되고 있다. [1]에서는 시간의 흐름에 따라 사람들이 관심을 갖는 주제를 파악하기 위해서 주요 구문을 찾는 시도를 하였으며, [2]에서는 주어진 키워드에 대한 동향을 분석하기 위해 특이값 분해(SVD)를 사용하였다. 하지만 [1]의 연구는 같은 구문이라도 다른 의미에서 사용 가능하다는 사실을 잘 반영하지 못하고, [2]는 사람이 직접 키워드를 넣어 주어야 하기 때문에 블로그 글들에서 주제를 찾기 어렵다. 본 연구에서 사용하는 LDA(Latent Dirichlet Allocation)는 한 단어가 갖는 여러 의미가 문맥에 따라 명확히 구분되며, 데이터 속에서 주제를 자동으로 발견해준다.[3] 기존에는 LDA를 주로 주제가 한정되어 있는 학술 문서에 사용하였는데 이 연구에서는 주제가 매우 광범위하고 사이즈가 큰 블로그 데이터에 적용해봄으로써 LDA의 장점들이 여전히 제 역할을 하는지, 어떤 문제가 생기는지 살펴볼 것이다.

본 논문의 전체 구성은 다음과 같다. 2장에서는 연구를 위해 사용된 유형별 데이터 셋에 관해 설명하며, 3장에서는 LDA와 평가 방법에 대해 살펴본다. 4장에서는 LDA 학습 결과를 통해 그 특징과 의미를 분석한 후, 5장에서 결론 및 향후 계획을 제시한다.

2. 데이터 소개

본 연구를 위한 기본 데이터로는 TREC BLOGS08 데이터 셋[2]이 사용되었다. BLOGS08 데이터 셋은 University of Glasgow에서 인터넷 상의 블로그들로부터 수집한 2천 8백만 개의 블로그 포스트들로 구성되어 있으며 블로그 연구를 위한 표준 데이터 셋을 제공할 목적으로 제작되었다.

BLOGS08 데이터 셋의 구성은 [표 1]과 같다. BlogSpot과 LiveJournal의 포스트가 전체 데이터의 55% 이상을 차지하고 있다. 이 연구에서는 전체 포스트 중 다음 조건을 만족하는 85,000개의 포스트를 선택하여 사용하였다.

- 50개 이상의 단어로 이루어져 있음.
- HTML에서 본문에 해당하는 영역을 쉽게 추출할 수 있음.
- 2008년 1월부터 2008년 12월 사이에 작성됨.

전체 포스트	BlogSpot	LiveJournal	기타
28,517,411(100%)	10,232,739(36%)	5,933,478(21%)	12,351,194(43%)

표 1 TREC BLOGS08 데이터 셋 통계

1) http://www.sifry.com/alerts/archives/000493.html

2) http://trec.nist.gov/data.html

de	time	la	people	en
day	don	post	el	postcount

표 2 제거된 단어 리스트

데이터종류	포스트 수	단어 종류	저자 수	포스트별 평균 단어 수
Original	85125	104448	5539	251.69
Refined	72143	53257	4844	225.24
Limited-10	23069	31343	4838	225.56
Limited-20	32544	36922	4841	225.48

표 3 분석 대상 데이터 셋

이것을 이용하여 다시 서로 다른 네 가지 유형의 데이터 셋을 만들었는데 각각의 특징은 다음과 같다.

- Original: 처음 선정된 데이터 그대로.
- Refined: Original 데이터 셋에서 영어가 아닌 언어로 쓰인 포스트들을 삭제하고, 빈도가 가장 높은 10개 단어 제거. [표 2]에 제거된 단어들이 나타나 있으며, 4.1에서는 이 단어들이 주제 선정에 미치는 영향에 대해 알아볼 것이다.
- Limited-10: Refined 데이터 셋에서 한 저자당 포스트 수를 최대 10개로 제한함.
- Limited-20: Refined 데이터 셋에서 한 저자당 포스트 수를 최대 20개로 제한함.

Limited 데이터 셋들은 저자당 포스트 수가 제한될 때 주제 분포가 어떻게 변하는지 확인하기 위해 사용되었다. [표 3]에 각 데이터 셋에 관한 통계들이 나타나있다.

기본적으로 모든 데이터 셋은 Porter Stemmer를 사용하여 비슷한 뜻을 지니는 단어들이 한 단어로 나타날 수 있게 하였으며, 전체 문서에 대해 stemming 과정이 끝난 후 7번 이상 나타난 단어들만을 분석 대상으로 선정하였다.

3. 방법론

3.1. LDA

LDA는 문서를 작성하는 과정에 관한 생성(Generative) 모델이며 문서들의 주제 분류를 위해 고안되었다. 각 문서 내에 여러 주제들이 혼재되어 있다는 것을 가정하며, 분석 결과를 새로운 문서에 적용하여 새로운 문서의 주제도 분석해낼 수 있다.

모델을 통해 주어진 문서의 주제를 분석하기 위해서는 추론(inference) 과정이 필요하다. 지금까지 추론을 위해 다양한 방법들이 제시되었다. [3]에서는 EM을 이용한 추론 방법을 제시했으며, [4]는 깁스 샘플링(Gibbs sampling)을 이용한 방법을 제시하였다. 깁스 샘플링은 국소 최대값(local maxima)에 갇히는 문제가 생기지 않으며 구현도 비교적 간단하다는 특징이 있다. 본 논문에서는 깁스 샘플링을 이용하여 추론하였다. 이러한 추론을 통해 모델에서 사용하는 매개변수의 값을 구하면, 추론에 사용된 문서들의 주제 분석이 가능함과 동시에 처음 보는 문서들의 주제도 분석해낼 수 있다.

3.2. 평가 방법

본 연구에서는 각각의 데이터 셋을 이용해 LDA 모델을 학습시킨 후 그 성능을 측정하기 위해 혼잡도(perplexity)의 개념을 도입하였다. 혼잡도는 학습된 생성 모델이 실제 관찰 가능한 결과를 생성해낼 확률을 측정하는 것으로, 모델의 혼잡도가 낮을수록 성능이 좋음을 의미한다. 실험 환경별로 각 모델의 혼잡도를 구하기 위해 그 모델을 학습시킬 때 사용했던 문서들 중 2000개의 문서를 임의로 추출하여 테스트 데이터로 사용했다.

혼잡도의 정확성을 위해 여러 번 샘플링 하여 그 값을 구하였다. 여기서의 샘플링은 모델의 매개변수들의 값을 구하는 전체 과정 한 번을 뜻하며, 여러 번의 샘플링을 통해 추론된 매개변수들을 사용하여 보다 정확한 혼잡도를 구할 수 있다.

4. 결과

4.1. 주제 분포

Refined 데이터 셋을 사용하여 LDA 모델을 학습 시켜 얻은 각 주제별로 확률이 가장 높은 10개의 단어들을 [표 4]에 나타냈다. 주제 41에는 은행과 관련된 단어들이 나타나 있다. 또한 주제 16, 22, 35, 37은 각각 문학, 인종, 라디오, 요리와 관련된 단어들로 구성되어 있다. 왼쪽 3개의 주제는 50개의 주제들 중 상위 3개 단어의 확률의 합이 가장 높은 주제들이다.

동일한 데이터 셋을 사용하여 주제 수를 50에서 200으로 늘렸을 때에는 기존에 존재하던 주제들은 대체로 유지되면서 매우 구체적인 주제가 새로 등장하는 것을 볼 수 있었는데, [표 5]에 나타난 주제들은 이전 50개 주제들에선 볼 수 없었다. 또한 주제의 수가 증가함에 따라 한 단어가 여러 주제에서 동시에 높은 확률로 나타나는 현상도 관찰할 수 있었는데, 134번 주제의

주제 41	
단어	확률
account	0.0601
review	0.0279
bank	0.0260
month	0.0251
cd	0.0203
rate	0.0188
check	0.0167
deposit	0.0165
credit	0.0155
save	0.0153

주제 16	
단어	확률
book	0.0279
read	0.0266
write	0.0194
writer	0.0114
stori	0.0113
author	0.0103
raglin	0.0098
publish	0.0094
reader	0.0087
poem	0.0077

주제 22	
단어	확률
black	0.0184
white	0.0157
american	0.0140
wright	0.0140
king	0.0120
race	0.0120
america	0.0115
call	0.0098
racist	0.0097
flag	0.0089

주제 35	
단어	확률
chicago	0.0229
public	0.0188
jeff	0.0162
affair	0.0135
week	0.0128
channel	0.0099
radio	0.0099
berkowitz	0.0093
watch	0.0089
illinoi	0.0089

주제 37	
단어	확률
food	0.0137
recip	0.0102
cook	0.0099
eat	0.0078
cup	0.0073
add	0.0065
tast	0.0064
bake	0.0064
minut	0.0063
egg	0.0062

표 4 Refined 데이터 셋으로 50개 주제를 학습한 결과

주제 102	
단어	확률
nuclear	0.008395
tower	0.007752
wtc	0.006991
nuke	0.006543
explos	0.005616
steel	0.005527
reactor	0.005109
bomb	0.004825
syndrom	0.003839
collaps	0.003480

주제 134	
단어	확률
plant	0.029777
food	0.025209
garden	0.020689
farm	0.016497
seed	0.009417
crop	0.008712
farmer	0.008607
gm	0.008231
feed	0.007832
grow	0.007796

표 5 주제 수를 200으로 늘린 결과

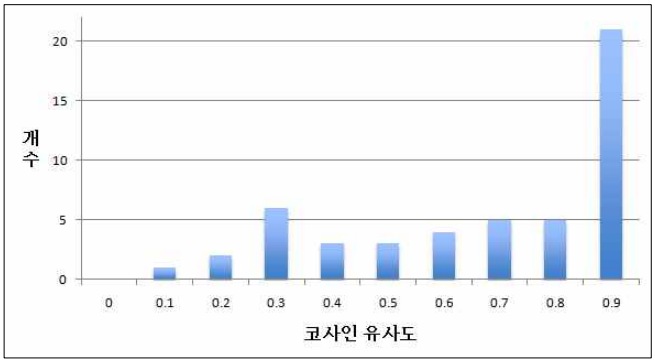

그림 1 샘플 1에 대한 샘플 2의 유사도 분포

단어	확률	단어	확률
don	0.026262	didn	0.010457
time	0.022427	lot	0.009178
peopl	0.019710	talk	0.008695
ve	0.019059	fill	0.008531
ll	0.011734	call	0.008460

표 6 빈도가 높은 단어들로 이루어진 의미 없는 주제

*food*는 주제를 50개로 제한했을 경우 *recipe*, *cook*, *eat* 등과 함께 요리와 관련된 37번 주제를 구성하던 단어였다. 이 단어는 200개의 주제를 선정했을 때 기존의 *recipe* 등과 함께 요리를 뜻하는 주제에서 주요 단어로 선정이 되었을 뿐 아니라 134번 주제와 같이 작물과 관련된 주제에서도 주요 단어로 선정이 되었다. 이를 통해 한 단어가 의미가 다른 여러 주제에서 동시에 주요한 역할을 할 수 있음을 알 수 있다.

한 가지 주제에 대한 글이 데이터 셋에 많이 들어있으면 LDA 학습 결과 하나의 주제로 나타날 가능성이 많아진다. [표 4]의 35번 주제에는 고유명사가 많이 나타나는데 확인 결과 이 단어들이 주를 이루는 한 블로그의 포스트들이 데이터 셋에 많이 포함되어 있었다. Limited-10 이나 Limited-20 데이터 셋처럼 한 저자로부터 나올 수 있는 포스트의 개수를 인위적으로 제한시키면, 많은 사람들이 관심을 가지고 글을 쓰는 주제들은 여전히 데이터 셋에서 높은 비율을 차지하지만 몇몇의 커다란 블로그들이 LDA 결과에 끼치는 영향은 줄어든다. Refiend 데이터 셋의 단어 빈도는 1419.15의 표준 편차를 가지고 있었던 반면 Limited-10 데이터 셋의 단어 빈도는 586.34의 표준 편차를 가지고 있었다. 다시 말해, 데이터 셋 내의 단어 빈도의 대비(contrast)가 낮아지면서 Refined 데이터 셋에서는 매우 낮은 비율을 차지했던 주제들이나 단어들이 Limited-10 이나 Limited-20 데이터 셋에서는 비중이 높아지는 효과를 얻는다. 더 자세한 분석은 4.2.2와 4.3.3에서 다루도록 한다.

일반적으로 여러 가지 주제에서 널리 자주 쓰이는 단어가 모델 학습에 어떤 영향을 주는지 알아보기 위해 Original과 Refined 데이터 셋을 통해 나온 결과를 비교하였다. Original 데이터 셋을 통해 모델을 학습 시킨 결과 하나의 주제를 이룬다고 생각할 수 없는 단어들로 구성된 주제가 항상 등장했다. [표 6]에 이 주제들의 주요 단어와 그 단어들의 확률이 나타나 있다. 이를 통해 빈도가 높은 모호한 단어들이 하나의 주제로 모여드는 현상을 관찰할 수 있었다. [표 6]과 [표 2]을 비교하면 많은 단어들이 중복되는 것을 알 수 있다. 또한 [표 2]에 나타난 많은 단어들이 주제와는 상관없이 여러 주제에서 동시에 나타나는 것을 확인할 수 있었다.

4.2. 주제 유사도

4.2.1 같은 조건에서의 샘플링 비교

각각의 샘플링이 언제나 비슷한 결과를 도출하는지 보기 위해 각 샘플링에서 나타나는 주제들의 유사도를 측정해보았다. 주제 간의 유사도를 수학적으로 측정하기 위해 두 주제를 이루는 단어들의 확률 분포 벡터의 코사인 유사도를 구하였다. 두 샘플링 결과 주제들의 유사도를 계산한 뒤, 첫 번째 샘플링의 각 주제와 유사도가 가장 높은 주제를 두 번째 샘플링 결과에서 찾아내어 그 분포를 [그림 1]에 나타냈다. 전체 50개의 주제들

코사인 유사도 = 0.999697			
샘플링 1 (주제 49)		샘플링 2 (주제 22)	
단어	확률	단어	확률
account	0.0764	account	0.0759
review	0.0621	review	0.0605
bank	0.0512	bank	0.0507
month	0.0330	cd	0.0325
cd	0.0313	month	0.0323
rate	0.0276	rate	0.0279
check	0.0235	check	0.0229
credit	0.0153	credit	0.0148
deposit	0.0147	deposit	0.0145
deal	0.0128	save	0.0128

코사인 유사도 = 0.524881			
샘플링 1 (주제 8)		샘플링 2 (주제 21)	
단어	확률	단어	확률
photo	0.0145	south	0.0154
tree	0.0137	north	0.0144
weather	0.0104	weather	0.0122
rain	0.0102	island	0.0116
garden	0.0101	rain	0.0104
bird	0.0088	storm	0.0100
snow	0.0084	snow	0.0096
wind	0.0079	west	0.0080
trip	0.0075	winter	0.0078
winter	0.0074	wind	0.0074

코사인 유사도 = 0.178105			
샘플링 1 (주제 13)		샘플링 2 (주제 6)	
단어	확률	단어	확률
chicago	0.0254	obama	0.0595
public	0.0180	vote	0.0268
jeff	0.0178	mccain	0.0249
affair	0.0151	democrat	0.0246
toni	0.0113	campaign	0.0193
rep	0.0108	senat	0.0170
berkowitz	0.0105	republican	0.0169
illinoi	0.0100	elect	0.0169
includ	0.0095	candid	0.0159
week	0.0092	clinton	0.0143

표 7 샘플링 1과 샘플링 2의 코사인 유사도

중 21개의 주제들이 0.9에서 1.0 사이의 유사도를 지니고 있다. [표 7]은 유사도별로 해당 주제의 상위 10개 단어를 나타낸 결과이다. 첫 번째 샘플링의 49번 주제와 두 번째 샘플링의 22번 주제는 상위 10개 단어 중 9개 단어가 동시에 나타나며, 0.52의 유사도를 가지는 8번 주제와 21번 주제는 5개의 공통 단어를 공유한다. 하지만 가장 낮은 유사도를 보여준 13번 주제와 6번 주제는 상위 10 단어 내에 공통되는 단어를 찾을 수 없다. 첫 번째 샘플링의 50개의 주제 중 절반 이상이 두 번째 샘플링의 주제들과 0.8 이상의 유사도를 가지며, 여러 샘플링 결과에 걸쳐 동시에 나타나지는 않는 주제들도 있었다.

4.2.2 다른 데이터 셋 비교

Refined와 Limit-10 데이터 셋의 결과에 대해 각각 자가 주제 유사도—하나의 샘플링 내에서 주제 쌍들의 코사인 유사도—를 계산한 뒤 가능한 모든 주제 쌍 중에서 유사도가 0.1보다 큰 쌍의 개수 비율을 보면 각각 0.1012와 0.0546이다. 즉 Refined 데이터 셋의 결과 주제들이 Limit-10 데이터 셋보다 서로 더 비슷하다는 뜻이다. 이는 데이터 셋 내의 단어들의 영향력 대비가 줄어든다는 것을 뒷받침 해준다.

4.3 혼잡도 측정

4.3.1 깁스 샘플링 횟수별

Refined-10 데이터 셋을 사용하여 깁스 샘플링 횟수가 증가함에 따라 혼잡도가 어떻게 변하는지 측정하였다. 샘플링은 총 10번 하였고, 결과는 [그림 2]의 왼쪽 그래프에 나와 있다. 깁스 샘플링이 한 번 끝난 뒤 혼잡도는 7,000이 넘었지만 20번이 지나고 나서는 절반 이상 줄어들었다. 그리고 점점 수렴하여 깁스 샘플링이 300번 이상 진행되면 혼잡도에 거의 변화가 없었다. 이때까지 소요되는 시간은 Intel Core2 Quad CPU 2.40GHz, 4GB 메모리를 사용했을 때 약 9시간 정도이다.

4.3.2 주제 개수별

주제 개수에 따른 혼잡도의 비교는 [그림 2]의 오른쪽 그래프에 나와 있다. 처음 세 개는 모두 Refined 데이터 셋이며 각각 50, 100, 200개의 주제로 학습하였다. 주제가 50개일 때 혼잡도가 가장 낮고 100개일 때 혼잡도가 가장 높았다.

4.3.3 제한된 포스트의 개수별

Limited-10, Limited-20 데이터 셋을 이용하여 혼잡도를 비교한 결과가 [그림 2]의 오른쪽 그래프에 나와 있다. 그래프의 가장 오른쪽에 있는 두 막대를 비교해보면 포스트의 개수를 10개로 제한했을 때 혼잡도가 20개로 제한했을 때보다 높고 이는 Refined 데이터 셋의 혼잡도보다 훨씬 높은 수치이다.

4.2.2에서 Limited-10 데이터 셋의 결과 주제들이 Refined 데이터 셋의 결과 주제들보다 더 명확하다고 했다. 그럼에도 불구하고 혼잡도는 높아지는 이유는 두 데이터 셋의 특징 자체에서 유추할 수 있다. 한 블로그의 모든 포스트들이 비슷한 형식을 갖는 경우가 있는데 특히 한 주제에 대해 전문적인 블로그들에

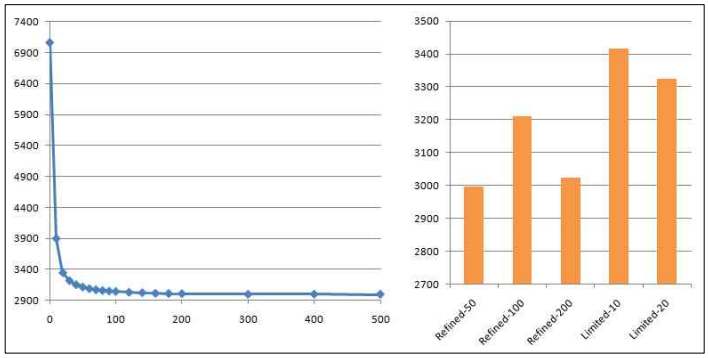

그림 2 샘플링 횟수와 주제 개수에 따른 혼잡도

서 그런 현상이 자주 나타난다. Refined 데이터 셋에 포함된 이런 포스트들이 전체적인 혼잡도를 크게 낮추는 역할을 해주는 것이다. 정확한 분석은 추후 연구로 남겨둔다.

5. 결론

LDA 문서 모델을 블로그 포스트들처럼 주제가 광범위하고 양이 많은 데이터에 적용시키기 위해 이 실험을 시작하였다. 그 결과 은행이나 문학과 같은 큰 범위의 주제들이 나타났다. 블로그마다 가지고 있는 포스트의 양이 다르므로 이것이 LDA 학습에 끼치는 영향을 알아보기 위해 하나의 블로그에서 데이터 셋에 포함시킬 포스트의 수를 제한시켰을 때, 그렇지 않을 때에 비해 단어 빈도 표준 편차가 매우 낮아지고 양이 비대한 블로그가 주제를 차지하는 현상을 제거할 수 있었다. 또한 혼잡도가 수렴할 때까지 깁스 샘플링을 하면 대용량 데이터를 처리하는 데에 큰 무리 없는 속도가 나오는 것을 확인하였다.

LDA가 자동으로 만들어내는 주제를 어느 정도 제어하기 위한 엄밀한 분석은 차후 연구로 남겨 둔다. 모델의 성능을 분석하기 위해 샘플링이 많이 필요한데 속도를 최적화하기 위한 연구도 가치 있을 것이다.

6. 감사의 글

본 연구는 한국과학기술원 정보통신대학의 2009년도 BK21사업 지원을 받아 수행되었습니다.

7. 참고 문헌

[1] N. Glance, M. Hurst, T. Tomokiyo, BlogPulse: Automated Trend Discovery for Weblogs, In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.

[2] Y. Chi, B. L. Tseng, J. Tatemura. Eigen-trend: Trend analysis in the blogosphere based on singular value decompositions, In Proceedings of the 15th CIKM Conference, 2006.

[3] D. M. Blei, A. Y. Ng, and M.I. Jordan, Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, 993-1022, 2003.

[4] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The Author-Topic Model for Authors and Documents, AUAI04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 487-494, 2004.