P2P 환경에서 확률적 잠재 의미 분석에 기반한 정보 검색

구태완, 김유섭, 이광모
한림대학교 컴퓨터공학과
e-mail : taewani@hallym.ac.kr

Information Retrieval based on Probabilistic Latent Semantic Analysis within P2P Environments

Tae-Wan Gu, Yu-Seop Kim, Kwang-Mo Lee
Dept. of Computer Engineering, Hallym University

요 약

전통적인 Peer-to-Peer 모델에서 정보검색 문제를 해결하기 위한 방법으로는 질의 및 키워드를 각 Peer 에 전송하여 해당 질의 및 키워드와 문서들을 비교하는 방법이 대부분이었다. 본 논문에서는 이러한 방법을 확장하여 문서에 대한 의미론적 분석을 통해 검색의 정확성을 향상시키고자 한다. 이를 위해 본 논문에서는 확률적 의미분석 기법을 이용하여 각 Peer 에 존재하는 정보에 대한 색인을 작성 한 후, 이것을 Peer-to-Peer 환경에 적용하기 위한 분산 색인 분배 알고리즘을 제안한다.

1. 서론

Peer-to-Peer(P2P) 기반의 네트워크 모델은 기존의 클라이언트/서버(Client/Sserver:C/S) 모델과는 달리 서비스의 주체가 되는 서버가 존재 하지 않기 때문에 서버에서 발생 하는 부하를 네트워크 말단으로 옮겨 성능을 향상 시킬 수 있다는 장점이 있다[1]. 그러나 독립적인 서버가 존재 하지 않기 때문에 분산되어 있는 노드들에서 특정한 정보를 검색하는 것은 P2P 네트워크 모델에서 해결해야 할 문제중의 하나이다[8].

최근 연구들[2, 3, 4]에 의하면 P2P 를 기반으로 하는 검색 기술들이 소개 되고 있지만 대부분은 분산 해싱 테이블(Distributed Hasing Table:DHT)을 이용하여 대상 노드를 빨리 검색하고, 질의어를 이용한 단순 매칭 기법을 적용하여 해당 노드에서 질의어와 매칭되는 문서를 검색하는 형식의 연구만 이루어져 왔다[5, 6]. 그리고 웹 검색의 경우 검색된 문서에 가중치를 적용하는 방법이 있다[7]. 그러나 이것은 질의어의 중의성 및 다의성 문제를 야기 시키게 된다. 이와 관련해 [8]은 잠재 의미 색인(Latent Semantic Indexing:LSI)을 P2P 환경에 적용한 연구이지만 검색의 정확성은 LSI 의 성능을 넘지 못한다[8].

본 논문에서는 LSI 에 비해 나은 성능을 가진 것으로 알려진 확률적 잠재 의미 분석(Probabilistic Latent Semantic Analysis:PLSA)[9]을 사용한다. 이것은 노드 정보와 질의간의 관계를 의미 공간(semantic space)에 벡터로 표현 한 후 그 유사도에 근거한 검색이 이루어 진다. PLSA 는 LSI 에 비해 더 나은 검색 성능을 보인다[9]. 하지만 PLSA 와 같이 의미론적 접근에서는 다중 차원의 벡터가 생성되므로 이것을 네트워크 모델에 적용하기 위한 논의가 필요하다[8]. 때문에 네트워크 모델과 벡터간의 차원 불일치 문제를 해결하기 위해 단순 네트워크 모델인 Pastry[6]를 기반으로 네트워크를 구성한다. 그리고 이를 기반으로 색인을 유지하고 관리하기 위한 분산 색인 분배 알고리즘을 제안하며, 실험을 통해 기존의 연구들에 비해 정확도와 시스템 자원 활용 및 부하에 대해 유용함을 보인다.

2. 확률적 잠재 의미 분석(Probabilistic Latent Semantic Analysis:PLSA)

확률적 잠재 의미 분석(PLSA)은 문서들에서 동시에 나타나는 관찰들이 잠재 클래스 변수(latent class variable) $z \in Z = \{z_1, z_2, \cdots, z_K\}$ 와 연관된 양상 모델(Aspect Model)에 기반한다[9]. 단어가 문서에 동시에 나타날 확률 $P(d,w)$는 식 (1)과 같으며 이때 d 는 문서, w 는 단어를 의미하고, $P(w|z)$ 와 $P(d|z)$ 는 토픽에 근거한 단어와 문서의 분포를 나타낸다.

$$P(d,w) = \sum_z P(z)P(d,w|z) = \sum_z P(z)P(w|z)P(d|z) \quad (1)$$

결국 위의 식 (1)에 대한 해를 찾기 위해 PLSA 에서는 우도함수(likelihood function) L, 즉 식 (2)를 최대화 하는 파라미터 $P(z)$, $P(d|z)$, 그리고 $P(d|z)$를 찾아야 한다. 이때 최대화는 EM (Expectation Maximization) 알고리즘을 이용하여 수행된다.

$$L = \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(d,w) \quad (2)$$

이렇게 각 문서의 출현 확률값을 구한 뒤, 각 문서의 유사관계를 정의한다. 다음의 식 (3)은 토픽 z 에 대한 문서 d_1과 d_2의 유사성을 정의한다. 그리고 식 (4)는 식 (3)을 계산하기 위한 파라미터로 사용된다.

$$Sim = \frac{\sum_k P(z_k|d_1)P(z_k|d_2)}{\sum_k P(z_k|d_1)P(z_k|d_1)\sum_k P(z_k|d_2)P(z_k|d_2)} \quad (3)$$

$$P(z_k|d) = \frac{P(z_k)P(d|z_k)}{\sum_l P(z_l)P(d|z_l)} \quad (4)$$

3. P2P 환경의 문서 검색

3.1 P2P 문서 검색 구조

문서 검색을 위한 P2P 구조는 인덱싱 모듈(Indexing Module), 질의 모듈(Query Module), 전송 모듈(Transmission Module)로 나뉜다.

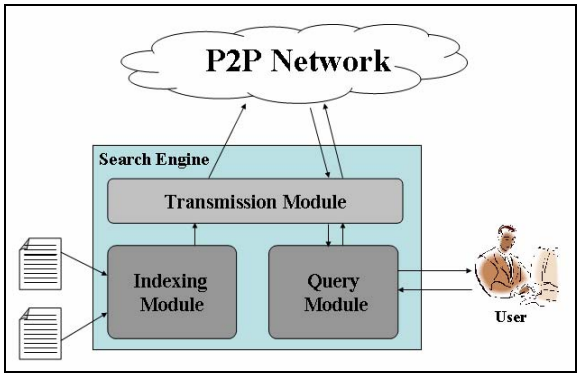
인덱싱 모듈(Indexing Module)은 PLSA 를 이용하여 인덱싱 과정을 수행하게 된다. 검색될 문서와 해당 문서에서 나타날 단어간의 확률을 계산한다. 그 후, 문서에 대한 문서-단어 벡터를 생성하며 이러한 벡터를 이용하여 실제 문서가 저장될 노드의 NodeID 를 계산한다. 이때 벡터와 NodeID 간를 매핑을 위한 매트릭스를 가지며, 이것은 네트워크 라우팅 테이블의 베이스에 의존적인 특성을 갖는다. 즉 네트워크 라우팅 테이블의 베이스가 낮다면 실제로 네트워크에 참여하는 노드의 개수 또한 적을것이므로 문서에 대한 벡터의 편차가 높아 질 수 있을 것이다. 인덱싱 모듈은 생성된 매트릭스를 기반으로 해당 문서를 전송하기 위해 전송 모듈을 호출하게 된다.

$$MAX\left[\frac{1}{k}\sum_k (P(d|z_k))^2 - \left(\frac{\sum_k P(d|z_k)}{k}\right)^2\right] \quad (5)$$

이때 문서가 전송될 노드의 위치 $Pos(d_i)$ 는 토픽 z 에 의해 결정된다. 그리고 식(5)를 이용해 확률 분포를 구성하고 그에 따른 확률값에 기반해 $Pos(d_i)$ 가 결정된다.

질의 모듈(Query Module)은 최초 질의가 도착 했을 때 사용자의 정보를 가지고 있으며, 인덱싱 모듈과 마찬가지로 PLSA 에 기반한 질의 확률값을 계산한다. 그리고 해당 질의를 전송 모듈로 전송하는 역할을 한다. 최종적으로 각 노드에서 전송되는 검색결과를 질의자(user)에게 전송하게 되며, 만약 검색이 실패했을 경우 이를 질의자에게 알리게 된다.

전송 모듈(Transmission Module)은 각 모듈로부터 받은 정보를 실제 네트워크 상의 노드로 전송하는 역할을 하며, 인덱싱 된 문서를 네트워크로 업로드 하는 기능과 검색 결과를 알려주는 메시징 기능이 있다. 다음의 [그림 1]은 본 논문에서 작성된 검색엔진의 구조를 나타낸다.



[그림 1] P2P 문서 검색 엔진

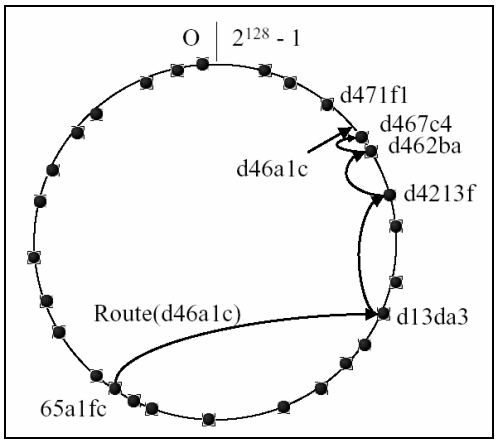
3.2 분산 색인 알고리즘

각 노드들은 자신의 고유한 의미공간을 가지게 된다. 이때 노드에 위치하는 의미공간을 식별하기 위해 네트워크 노드 식별자인 nodeID 를 해당 의미공간에 매핑한다. Pastry 는 모든 노드가 유일한 128 비트 nodeID 를 가진다[6]. 그리고 메시지 전송을 위해 라우팅되는 정보는 각 노드에서 유지하는 라우팅 테이블을 기반으로 하는데 라우팅 테이블은 b 개의 베이스로 이루어진다.

[그림 2]는 Pastry 에서 nodeID *65a1fc* 에서 *d46a1c* 로 메시지가 라우팅되는 그림인데, 각 노드의 nodeID 가 b 에 의해 정의되므로 이를 기준으로 트리형 구조로 표현할 수 있게 된다. 그러므로

검색되는 모든 문서들은 이 의미공간의 루트에 인접한 벡터로 표현되며, 질의도 함께 벡터로 표현되어 루트 노드를 기준으로 인접 노드로 확장하여 검색이 이루어진다.

이것은 생성되는 문서가 다른 의미 공간에 표현되어야 할 경우 정보의 전송을 줄일 수 있는 장점이 있으며 기본 알고리즘은 다음과 같다.



[그림 2] Pastry 에서 메시지 라우팅

(1) 새로운 문서가 생성되었을 때 문서의 벡터 V_d 를 인덱싱 모듈을 이용하여 생성하고 V_d 를 이용하여 네트워크 상의 노드로 전송하기 위해 NodeID 값을 계산한다. 벡터 V_d 를 갖는 문서는 앞에서 계산된 NodeID 값을 갖는 노드로 전송된다. 이때, 각 노드는 문서의 벡터와 NodeID 를 매핑하는 매트릭스는 검색 엔진의 인덱싱 모듈을 이용한다.

(2) 새로운 질의가 도착하였을 경우 질의의 벡터 V_q 를 질의 모듈을 이용하여 생성하고 V_q 에 대해서도 (1)의 과정을 반복한다.

(3) (2)의 과정을 거쳐 질의가 전송되면, 질의를 받은 노드는 질의와 문서간의 유사도를 측정하게 된다. 그런 다음 그 결과를 검색 엔진의 질의 모듈에게 반환한다. 하지만 로컬검색에서 원하는 문서를 발견하지 못하였을 경우 해당 노드의 자식 노드로 질의를 재전송한 후, 로컬 검색을 다시 수행한다.

(4) 최종 단말 노드에서도 검색이 이루어 지지 않으면 단말 노드는 질의 모듈로 "검색실패(Search Fail)" 메시지를 전송하고 검색을 마친다.

위의 과정에서 검색이 수행되는 동안 오로지 질의와 검색 결과만 전송된다. 그런데 전송되는 질의와 검색 결과는 크기가 매우 작고 검색 대상문서(corpus) 크기에 독립적이다. 그러므로 제안된 알고리즘은 PLSA 와 같은 검색 성능을 가지게 된다. 그리고 메시지 전송에 따른 비용은 네트워크에 참여하는 노드의 수를 N 이라고 할 때 $O(\log_{2^b} N)$ 으로 계산되며 이때 N 은 참여 노드의 수, b 는 라우팅 테이블의 베이스이다[6].

4. 실험

4.1 실험 환경

본 논문에서 실험을 위해 사용되는 검색 대상 문서(corpus)는 1988 년 TREC-7 문서 집합 중 일부인 AP 뉴스 문서 집합을 사용하였으며 문서의 개수는 79,919 개이며, 사용단어는 19,286 개이다. 모든 문서는 불용어(stopword)를 제거하고 이를 다시 스태밍(stemming)하여 사용하도록 한다. 또한 네트워크 구성을 위해 16 대의 노드를 구성하였으며 각 노드의 연결은 10MB 의 대역폭을 가진다.

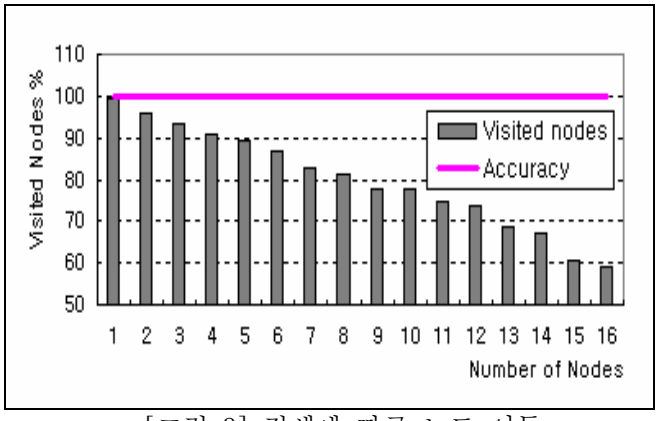
4.2 실험 결과

[표 1]은 인덱싱 모듈에 의해 전송된 문서들의 분포를 의미한다. 인덱싱 모듈에서 노드-문서간 매핑이 대체로 균등하게 분포됨을 알 수 있다. 이것은 전체 문서를 대상으로 검색이 이루어지는 것보다 해당 노드에서 보유하고 있는 문서에 대해서만 검색을 할 수 있으므로 검색 성능은 더 좋아 질 수 있다.

NodeID	문서 수	NodeID	문서 수
a111	2733	*b111*	4187
a112	2715	*b112*	4172
a113	5802	*b113*	4200
a114	6394	*b114*	4198
a211	3929	*b211*	7072
a212	4093	*b212*	6975
a213	4215	*b214*	7664
a214	4212	*b214*	5358

[표 1] 노드간 문서 분포

다음 [그림 3]은 문서 인덱싱 후, 질의에 대해 문서 검색을 위한 노드 이동과 검색 정확도를 나타낸다. [그림 3]에서 알 수 있듯이 전체 문서에 대해 검색을 하지 않고도 문서의 검색은 이루어 질 수 있으며 노드의 개수가 많을수록 노드 방문 횟수가 줄어 들어 노드 증가에 따른 불필요한 트래픽을 염려하지 않아도 된다.



[그림 3] 검색에 따른 노드 이동

5. 결론

본 논문은 P2P 환경에서 문서 검색을 위해 PLSA(Probabilistic Latent Semantic Analysis)를 이용한다. PLSA 는 문서에 대한 토픽을 정의하고 이 값을 이용하여 문서의 의미에 근거한 검색을 수행하는데, 이것을 P2P 네트워크에 적용하는데 따른 문제점을 해결하고 있다. 우선 의미 기반 검색에서 발생하는 차원 불일치 문제를 해결하기 위해 Ring 기반의 단일 차원 모델로 각 문서들을 매핑하고 있다. 이때 전체 문서벡터가 갖는 토픽 확률의 밀도가 높은 값을 기준으로 문서벡터를 분배하기 때문에 차원 불일치에 따른 차원 일치 연산을 수행하지 않아도 된다. 또한 이와 같은 방법으로 질의를 전송하기 때문에 전체 문서에 대한 검색을 필요로 하지 않는다. 하지만 질의를 받은 노드에서 질의 결과를 찾지 못할 경우, 다른 노드로 이동하여 검색을 수행하므로 검색에 있어 불필요한 트래픽 발생이 기대되었으나, 실제 검색 시 노드의 이동 횟수는 노드 수가 늘어 남에 따라 줄어들므로 이 문제는 검색 성능에 있어 큰 문제가 되지 않는다. 게다가 LSA(Latent Semantic Analysis)에 비해 나은 검색 성능을 갖는 PLSA 사용하므로 기존 연구에 비해 검색성능은 더 나을 것으로 예상된다.

참고문헌

[1] Dreamtech Software Team, *Cracking the Code Peer-to-Peer Application Development*

[2] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker., “Search and Replication in Unstructured Peer-to-Peer Networks”, *In ICS’02, June 2002.*

[3] E. Cohen, A. Fiat, and H. Kaplan., “Associative Search in Peer to Peer Networks: Harnessing Latent Semantics”., *In IEEE INFOCOM’03, April 2003.*

[4] K. Sripanidkulchai, B. Maggs, and H. Zhang., “Enabling Efficient Content Location and Retrieval in Peer-to-Peer Systems by Exploiting Locality in Interests”, *ACM SIGCOMM Computer Communication Review*, 32(1), January 2002.

[5] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek and Hari Balakrishnan, “Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications”, *SIGCOMM’01, August 27-31, 2001.*

[6] A. Rowstron and P. Druschel, "Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems", *Middleware'2001, Germany, November 2001*.

[7] http://www.google.com

[8] Chunqiang Tang, Zhichen Xu, Sandhya Dwarkadas, *Peer-to-Peer Information Retrieval Using Self-Organizing Semantic Overlay Networks*, *SIGCOMM’03,* August 25–29, 2003.

[9] Hofmann, T., “*Probabilistic latent semantic indexing,*” Proceedings of the 22th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR99), pp. 50.57, 1999.