LSI에서 질의 확장을 이용한 실험

안 성수, 김 동주, 이 기영, 김 한우
{ssahn, djkim, kylee, kimhw}@cse.hanyang.ac.kr
한양대학교 전자계산학과

Experiments using query expansion in LSI

Sung-soo Ahn, Dong-joo Kim, Ki-young Lee, Han-woo Kim
Department of Computer Science and Engineering, Hanyang University

요 약

한번의 질의로 사용자가 모든 요구를 표현하기 어렵고 만족시킬 수 없기때문에 질의를 확장하는 연구가 계속되고 있다. 본 논문에서는 LSI(Latent Semantic Indexing)에서 사용자의 질의와 의미공간에서의 용어들간의 유사도를 구해 최상위의 용어들을 순서를 정해 질의확장을 하는 방법과 LCA(Local Context Analysis)을 이용하는 방법을 제안한다. 그리고 문서 집합에 대해 3가지 가중치를 적용한 결과를 분석하고 질의확장시의 문제점과 향후 연구과제에 대해 설명한다.

1 서 론

컴퓨터가 생활의 필수가 되면서 사용자가 접하게 되는 정보는 과거에 비해 기하급수적으로 늘어가고 있다. 또한 인터넷과 컴퓨터의 발달로 사용자가 정보를 쉽게 찾을 수 있도록 검색엔진이 개발되고 있으며 그 기초가 되는 여러가지 방법들로 불리언 모델, 확률 모델, 벡터 모델등이 있다[1]. 본 논문에서 다루는 LSI는 벡터 모델의 연장으로 단어간의 비교로 검색이 실패하는 부분을 보완하는 모델이다[2]. 즉, 사용자의 질의에 있는 단어만을 검색하는 것이 아니라 같이 쓰이는 단어의 사용 패턴을 통계학적인 방법으로 찾아내는 방법이다. 본 논문에서는 LSI의 가장 큰 장점인 단어들이 유사하게 사용되면 같은 공간에 분포한다는 가정에 바탕을 두고 좀 더 성능을 향상시킬 목적으로 3가지의 가중치를 주고 LSI에 질의확장 실험을 하며 그 결과를 서로 비교하고 그 대안을 제시하고자 한다.

2 관련연구

사용자가 한번에 정보요구를 표현할 수 없기 때문에 검색된 문서중에 관련된 문서를 바탕으로 재질의하거나 질의의 용어들에 대해 가중치를 조정해서 다시 질의하는 방법들이 연구되어 왔다. 시소러스를 사용하는 방법으로 WordNet을 이용한 TREC데이터에 대한 실험이 있었지만 좋은 효과는 없었다[3]. 다른 방법으로 용어간의 근접성을 이용하고 어구에 가중치를 더 주는 방법[4]등이 있고 또한 질의에 의해 검색된 상위문서들이 관련있다는 가정하에 이들 문서들을 분석해 질의를 확장하는 방법[5]도 있다. LSI를 이용한 질의확장에 관한 연구는 임계치를 이용한 실험이 있었다[6].

3 Latent Semantic Indexing

LSI는 사용자의 문서정보에 대한 접근을 향상시키기 위한 방법이다. 대부분의 정보검색시스템은 데이터베이스의 단어와 사용자의 질의에 있는 단어간의 비교에 의해 문서를 찾는다. 사용자는 관심있는 토픽에 대해 매우 다양한 단어로 이를 표현하기 때문에 단어간의 비교는 완전히 정확하지는 않다. 즉, 같은 단어로 다른 객체들을 표현할 수 있기 때문에 부적합한 문서를 검색할 수 있다. (예를 들어 '배'는 문맥에 따라 다른 의미를 표현한다.) 역으로, 각각의 저자들이 같은 의미를 표현하기 위해 다른 단어들을 사용할 수 있기 때문에 관련된 문서가 검색되지 않을 수 도 있다. (예를 들어 '버스', '승용차', '택시' 등은 '자동차'에 대한 문서를 검색하려는 사용자에게 관심의 대상이 될 수 있다.) LSI는 이러한 문제를 해결하기 위해 용어와 문서간의 관계를 통계적인 방법으로 모델링하고 정보검색에 더욱 적합한 의미공간을 자동적으로 만든다. LSI는 일반적인 단어비교보다 몇가지 장점을 제공한다. 첫째, LSI는 사용자의 질의를 공유하지 않는 문서가 검색될 수 있도록 하고 30%의 성능향상을 보이고 있다. 둘째, 사용자의 질의와 비교해 순서화된 문서를 보여준다. 셋째, 용어와 문서가 LSI공간에 같이 표현되기 때문에 문서와 용어의 어떠한 조합도 질의로 사용될 수 있다.

4 실 험

질의를 확장하는 이유중의 하나는 사용자가 작성하는 질의의 단어와 문서가 인덱싱될 때의 단어가 같은 내용이라도 다르게 표현될 수 있기 때문에 이를 극복하려는 것이다. 본 논문에서는 LSI를 이용하면 같이 사용되는 단어들은 의미공간에서 같이 분포된다는 사실에 착안하고 질의확장을 실험한다.

4.1 실험환경

문서 집합	ADI	CISI	TIME
문서 갯수	82	1460	425
용어 갯수	378	5567	10884
문서당 평균용어	16	45	190
용어당 평균문서	4	13	8
질의당 평균관련문서	5	50	4
질의당 평균용어	5	8	8

표 1: Collection set의 특징

실험에 사용된 문서집합은 ADI, CISI, TIME이고 표 1과 같은 특징을 가지고 있다. Stopword는 SMART에서 사용된 것을 사용했고 실행환경은 Intel Pentium MMX 200, Linux kernel 2.0.35, 메모리 64MB 의 PC에서 진행되었다.

4.2 실험방법

가중치로 tf, tf·idf, log·entropy를 사용했으며 행렬을 근사화하는 k값은 ADI는 60, CISI, TIME은 100으로 정한 후 실험하였다[7]. 질의를 확장하는 방법으로 LSI 공간상에 용어를 표현하여 질의를 이 의미 공간에 투사해 용어와 질의간의 유사도를 코사인으로 측정해 확장할 용어의 순위를 정하였다. 질의는 원래의 질의, 그리고 최상위에 검색된 관련된 문서 즉 관련피드백, 그리고 질의를 LSI공간에서 10개, 20개, ...100개를 적용하였다. 최상의 관련문서를 질의로 준 것은 원래의 질의에 확장할 용어들을 더했을 때 최대로 얻을 수 있는 정확률로 생각했기 때문이다. 위의 확장된 질의에 대해 정확률, 재현율을 측정하였다. 다음장에 있는 표 2는 각각의 가중치에 대한 실험결과이다. 수치는 재현율 10%, 20%, ... 100%일 때의 평균정확률을 나타내고 있다.

LCA를 이용한 실험은 먼저 초기 질의를 주었을 때 최상위 5개의 문서가 관련있다고 가정하고 이 문서에 대해 local context analysis를 적용하였다. 최상위 5문서에 대해서만 적용한 이유는 각 질의당 문서집합에서 관련문서가 4개 또는 5개 정도이기 때문이고 더 많은 문서를 포함시킬 경우는 관련없는 문서를 분석해 더 나쁜결과가 산출되기 때문이다. 질의와 문서내에 있는 용어들의 유사도를 구했는데 이때의 기준 역시 SVD한 의미공간에서 코사인으로 랭

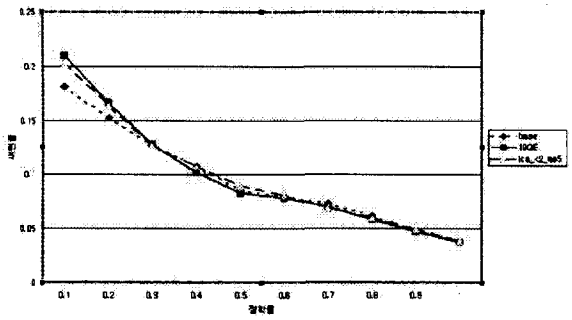

그림 1: CISI의 질의확장

킹을 정해 5개, 10개, 15개의 질의를 확장하였다. 사용한 k는 2,3,4,5를 적용하였다.

4.3 실험결과

먼저 LSI공간에서의 실험결과를 보면 용어의 가중치를 tf만을 주고 질의 확장을 했을경우 ADI와 TIME은 성능향상이 없었고 CISI는 10개 확장했을 때 몇 5.85% 향상되었고 가장 좋은 결과를 나타냈다. tf·idf의 경우 ADI는 100개를 확장했을 경우 7.36%, CISI는 10개 확장했을 경우 3.35%, TIME의 경우 100개 확장했을 경우 6.45%향상되고 가장 좋은 효과를 나타냈다. 그런데 가중치를 log·entropy를 주고 질의 확장을 하면 3개의 문서집합에 대해 원래의 경우보다 더 나빠짐을 나타내고 있다. 그림 1은 CISI에 대해 원래 질의, LSI공간에서 10개 확장한 질의 그리고 LCA를 이용한 것을 나타내고 있다.

LCA를 이용한 실험결과는 아래의 표 3과 같다. ADI문서는 3가지 가중치에 대해 향상이 없었고

문서 집합	ADI	CISI	TIME
tf	X[1]	X	0.25%
tf·idf	X	3.75%	1.85%
log·entropy	X	2.44%	2.53%

표 3: LCA를 적용했을 때의 평균정확률

CISI문서는 tf·idf의 경우 k=2, 5개의 용어를 확장시 3.75%, log·entropy의 경우 k=5, 5개의 용어를 확장시 2.44%향상되었다. TIME문서는 tf의 경우 k=2, 5개 용어를 확장시 0.25%, tf·idf의 경우 k=5, 5개의 확장시 1.85%, log·entropy의 경우 k=5, 5개의 용어를 확장시 2.53%향상되었다.

위의 결과를 보면 LCA를 적용하지 않고 가중치를 log·entropy로 주었을 경우 3가지 문서집합에 대해 향상이 없었는데 LCA를 적용시 2 문서집합에 대

[1] 평균 정확률이 향상되지 않았음을 뜻함

문서집합	ADI			CISI			TIME		
가중치	tf	tf·idf	log·entropy	tf	tf·idf	log·entropy	tf	tf·idf	log·entropy
원래질의	0.2493	0.2139	0.3061	0.0998	0.0955	0.1556	0.4456	0.3781	0.5445
RF	0.4979	0.4836	0.5329	0.1444	0.1332	0.2000	0.7491	0.7168	0.7314
10개 QE	0.2374	0.2140	0.2747	0.1056	0.0982	0.1429	0.4393	0.3815	0.5101
20개 QE	0.2342	0.2169	0.2730	0.1023	0.0969	0.1383	0.4390	0.3756	0.5087
30개 QE	0.2317	0.2140	0.2581	0.1015	0.0966	0.1355	0.4386	0.3761	0.4874
40개 QE	0.2311	0.2014	0.2484	0.1018	0.0968	0.1328	0.4369	0.3828	0.4820
50개 QE	0.2327	0.2050	0.2676	0.1021	0.0969	0.1309	0.4241	0.3840	0.4768
60개 QE	0.2359	0.2144	0.2681	0.1018	0.0963	0.1311	0.4259	0.3848	0.4809
70개 QE	0.2344	0.2198	0.2586	0.1021	0.0964	0.1310	0.4314	0.3929	0.4827
80개 QE	0.2277	0.2226	0.2442	0.1029	0.0970	0.1315	0.4293	0.3953	0.4794
90개 QE	0.2253	0.2249	0.2358	0.1022	0.0970	0.1317	0.4294	0.3968	0.4852
100개 QE	0.2216	0.2297	0.2304	0.1023	0.0962	0.1296	0.4323	0.4024	0.4943

표 2: 문서집합에 질의확장을 적용했을 때의 평균정확률

해 약 2.5%의 정확률이 향상되었다. 이는 entropy의 전역가중치를 줄 때 많은 문서와 이러한 문서에 포함된 용어들의 분포보다는 관련있다고 생각되는 문서들중의 용어를 분석해 entropy의 가중치를 주고 관련 용어를 선택하는 것이 더욱 효과적임을 나타내고 있다. 유사도를 구하는 계산시간과 처리면에서 LCA를 이용한 두번째 방법이 더욱 효율적임을 알 수 있다. 특이 주목할 것은 정확률이 가장 높은 log·entropy에서 향상이 있는 것이다.

5 결론 및 연구과제

LSI는 같이 사용되는 단어를 수학적인 모델링과 통계적인 처리를 이용해 기존의 벡터모델보다 성능향상을 가져온 기법이다. 이 모델을 이용해 좀 더 성능을 높이기 위한 방법으로 사용자의 질의와 문서집합내의 용어간의 유사도를 구하는 방법과 최상위 5개의 문서집합에서의 용어간의 유사도를 구해 질의확장의 실험을 하였다. 이는 LSI를 이용해 질의에 질의의 분포가 유사한 단어를 더함으로써 다시 한번 질의를 강조한 결과가 된다. 전체문서보다는 LCA를 이용한 방법이 계산상의 효율과 평균정확률에서 효과적임을 실험을 통해 알 수 있었다.

연구과제로 현재의 방법은 단지 원래의 질의에 용어를 더하기만 했는데 초기 질의에 관련되어 있지 않다고 생각되는 용어를 제거하는 방법, 확장할 용어 사이에 가중치를 다르게 주는 방법 그리고 Rocchio feedback과 같은 기법을 적용해서 원래의 질의에 가중치를 더 많이 주는 실험을 하는 것이다. 또한 LSI를 한글 검색 문서집합에 적용해 그 결과를 평가하고 한글의 특성에 맞는 질의확장 방법을 찾는 것이다.

참고 문헌

[1] Michael W. Berry, Murray Browne, "Understanding Search Engines, Mathematical Modeling and Text Retrieval," *Society for Industrial and Applied Mathematics, Philadelphia*, 1999

[2] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, 41(6):391-407,1990.

[3] Ellen M. Voorhees, "Query expansion using lexical-semantic relations," *Proc. of the 17th International Conference on Research and development in information retrieval*, pp 351-357, 1994

[4] Mandar Mitra, Amit Singhal, Chris Buckley, "Improving Automatic Query Expansioni," *Proc. of the 21th International Conference on Research and development in information retrieval*, pp 206-214, 1998

[5] Jinxi Xu, W. Bruce Croft, "Query expansion using local and global document analysis," *Proc. of the 19th International Conference on Research and development in information retrieval*, pp 4-11, 1996

[6] 임재현, 배희진, 김영찬. "용어 분포에 기반한 지능적 정보검색," 정보과학회논문지(B), 제 25권 제4호, pp 707-713, 1998.4

[7] Susan T. Dumais, "Improving the retrieval of information from external sources," *Behavior Research Methods, Instruments and Computers*, 23(2):229-236,1991