포스트 연관 점수의 분포 특성을 고려한 블로그 피드 검색 방법

송우상○, 이예하, 이종혁
포항공과대학교 컴퓨터공학과
woosang@postech.ac.kr, sion@postech.ac.kr, jhlee@postech.ac.kr

Using The Distributional Characteristics of Blog Post Relevance Scores for Blog Feed Search

Woosang Song○, Yeha Lee, Jong-Hyeok Lee
POSTECH, Department of Computer Science and Engineering

1. 서론

블로그 피드 검색(blog feed search, blog distillation)은 질의로 주어진 주제와 연관된 블로그를 찾는 것을 목적으로 한다. 블로그 피드 검색이 처음 소개된 2007년 Text REtrieval Conference(TREC) 이후 질의에 대한 블로그의 연관 점수 계산을 위한 다양한 방법들이 제안되었다[1-5]. 이 중에서 블로그를 구성하는 전체 포스트들의 평균 연관 점수를 사용하는 방법[3]이 널리 사용되고 있다. 이 방법은 질의에 대한 블로그의 전체적인 성향을 랭킹에 반영한다. 하지만 블로그 내의 포스트들이 나타낼 수 있는 평균 이외의 통계적 특성들은 랭킹에 반영되지 않았다.

본 논문에서는 [6]에서 제안된 Moment-based Ranking Model을 적용한 블로그 피드 검색 방법을 제안한다. 제안하는 방법은 블로그 포스트들의 연관 점수의 평균값 이외에 분산, 비대칭도(skewness), 첨예도(kurtosis) 등의 다양한 통계적 특성을 랭킹에 반영한다. 이를 통해 블로그 랭킹 과정에서 블로그 내의 전체 포스트들이 질의에 대해 나타내는 일관성을 랭킹에 반영할 수 있으며, 매개변수(parameter)를 통해 이에 대한 선호 정도를 조정할 수 있다.

2. 본론

Moment-based Ranking Model[6]은 랭킹 대상의 매개변수 추정치가 가지는 불확실성(uncertainty, risk)을 랭킹에 반영하기 위해 제안된 모델이다. 질의와 문서간의 연관성(relevance) 또는 문서의 언어 모델(document language model) 등이 랭킹과 관련된 매개변수가 될 수 있다. 본 논문에서 제안하는 블로그 피드 검색 방법은 질의에 대한 블로그의 연관성을 매개변수로 설정하여 Moment-based Ranking Model을 적용한다. 이를 통해 블로그의 연관 점수는 다음과 같이 계산된다.

$$\tilde{\theta} = -(1/b)\sum_{n=1}^{\infty}\kappa_n\frac{(-b)^n}{n!} = \mu - b\kappa_2/2 + \kappa_3 b^2/6 - \kappa_4 b^3/24 + \ldots \quad (1)$$

$\tilde{\theta}$는 매개변수 θ에 대한 추정값으로 해당 블로그에 대한 연관 점수로 사용된다. κ_n은 θ에 대한 n차 cumulant로써 각각 $\kappa_1=\mu$, $\kappa_2=\mu_2$, $\kappa_3=\mu_3$, $\kappa_4=\mu_4-3\mu^2$의 형태로 계산되며 κ_2는 분산, κ_3는 비대칭도, κ_4는 첨예도를 반영한다. b는 κ_n의 반영 정도를 조정하는 매개변수로 사용된다. θ의 평균 μ와 n차 central moment μ_n은 sample mean과 sample central moment 방식으로 다음과 같이 계산된다.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} Score(Q,D_i) \quad (2)$$

$$\mu_n = \frac{1}{N}\sum_{i=1}^{N}(Score(Q,D_i)-\mu)^2 \quad (3)$$

N은 블로그를 구성하는 포스트의 개수를 나타낸다. $Score(Q,D_i)$는 질의 Q에 대한 포스트 D_i의 연관 점수를 나타내며 언어 모델 기반 검색 방법인 Query Likelihood Model을 통해 계산된다. 포스트 D_i의 언어 모델은 Maximum likelihood estimation과 Dirichlet prior smoothing[7]을 통해 추정되었다. 블로그 랭킹 방식은 매개변수 b값의 조정에 따라 다음의 세 가지 유형으로 분류될 수 있다.

① Risk-ignore ($b=0$) : 질의에 대한 블로그의 연관 점수로 블로그를 구성하는 전체 포스트들의 평균 연관 점수만을 사용한다. 이는 [3]에서 제안된 것과 동일한 방법이다.

② Risk-averse ($b>0$) : 포스트들의 평균 연관 점수가 같을 경우 포스트들의 개별 연관 점수가 평균에서 크게 벗어나지 않는 블로그를 선호한다. 이는 질의로 주어진 주제에 대한 전체 포스트들의 일관성을 중요시하는 방식이다.

③ Risk-reward ($b<0$) : 포스트들의 평균 연관 점수가 같을 경우 질의에 대한 일관성이 낮은 블로그를 선호한다.

즉, 연관 점수 기준 상위의 포스트들이 높은 연관 점수를 가지는 블로그를 선호하는 방식이다.

본 논문에서 제안한 방법을 실험하기 위해 2007년과 2008년 TREC 블로그 피드 검색 task에 적용된 BLOGS06 COLLECTION[1]과 질의 및 평가 데이터가 사용되었다. 성능 평가 척도로는 MAP(Mean Average Precision), P@10(Precision at 10), NDCG(Normalized Discounted Cumulative Gain)[8]를 사용하였다.

표1. 2007, 2008년 질의 기준 블로그 피드 검색 성능 비교

	2007년 질의			2008년 질의		
	MAP	P@10	nDCG	MAP	P@10	nDCG
BASELINE	0.3339	0.5156	0.5554	0.2350	0.3960	0.4511
MOMENT-2	0.3447 (+4.13%)	**0.5200 (+0.85%)**	0.5718 (+2.95%)	0.2477 (+5.40%)	0.4120 (+4.04%)	0.4719 (+4.61%)
MOMENT-4	**0.3513 (+5.21%)**	0.5178 (+0.42%)	**0.5767 (+3.84%)**	**0.2512 (+6.89%)**	**0.4140 (+4.55%)**	**0.4794 (+6.27%)**

표 1은 매개변수 b가 최적으로 설정되었을 때 2007년 질의 45개와 2008년 질의 50개에 대한 각 검색 모델의 성능을 나타낸다. BASELINE은 risk-ignore 방식이며, MOMENT-2와 MOMENT-4는 수식 (1)에서 각각 2번째 항과 4번째 항까지를 사용한 검색 방법이다. MOMENT-2와 MOMENT-4 방식은 MAP을 기준으로 2007년 질의에서 BASELINE 대비 4~5%, 2008년 질의에서 5~7%의 성능 향상을 기록하였다. 또한 P@10과 nDCG에 대해서도 성능의 향상을 기록하였다. 이를 통해 포스트들의 연관 점수에 대한 분산 등의 다양한 통계 특성을 랭킹에 반영할 경우 포스트들의 평균 연관 점수만을 사용한 경우보다 검색 성능을 향상시킬 수 있음을 보여준다. 매개변수 b값은 2007년 질의의 P@10을 제외한 모든 경우에서 [-0.7,-0.5] 구간 내에서 설정될 경우 최적의 성능을 보여주었다. 이를 통해 risk-reward 방식이 risk-averse, risk-ignore 방식보다 블로그 피드 검색에 효과적이라는 것을 알 수 있다. 일반적으로 블로그는 블로그 저자의 다양한 관심사를 반영하기 때문에 질의와 연관된 블로그라 할지라도 주제와 연관되지 않는 포스트를 다수 포함하는 경우가 많다. Risk-reward 방식은 이러한 블로그에 대해 risk-ignore, risk-averse 방식에 비해 높은 연관 점수를 할당하기 때문에 블로그 피드 검색 성능을 향상시킬 수 있다. 이를 통해 블로그 피드 검색에서 블로그 내의 전체 포스트를 고려한 일관성 보다는 질의와의 연관성이 높은 상위의 포스트들의 연관 점수가 블로그 피드 검색 성능을 향상시킬 수 있는 중요한 요소임을 알 수 있다.

3. 결론

본 논문에서는 블로그 내의 포스트들의 평균 연관 점수 외에 분산, 비대칭도, 첨예도 등의 다양한 통계 수치를 반영한 블로그 피드 검색 방법을 제안하였다. 이를 통해 블로그 내의 포스트들이 질의에 대해 나타내는 성향을 더욱 정확하게 반영할 수 있었으며, 매개변수의 조정을 통해 블로그가 질의에 대해 나타내는 일관성에 대한 선호 정도를 조정할 수 있었다. 실험 결과 제안된 방법은 기존의 블로그 피드 검색 방법에 비해 MAP을 비롯한 검색 성능을 향상시키는 결과를 보여주었다. 앞으로의 연구에서는 제안된 모델을 기반으로 블로그 피드 검색 성능을 더욱 향상시킬 수 있는 방법과 매개변수의 자동 최적 설정에 대한 연구가 이루어져야 할 것이다.

감사의 글

본 논문은 2010년도 두뇌한국21사업, 포항공과대학교 정보통신연구소 자체 학술연구과제(선도과제), 그리고 한국과학재단 기초연구사업(No. 2010-0012662)의 지원으로 수행되었습니다.

참고문헌

[1] C. Macdonald, I. Ounis, and I. Soboroff, "Overview of TREC-2007 Blog track," in *Proc. of TREC-2007*, 2008.
[2] I. Ounis, C. Macdonald, and I. Soboroff, "Overview of TREC-2008 Blog track," in *Proc. of TREC-2008*, 2009.
[3] J. Arguello, J.L. Elsas, J. Callan, J.G. Carbonell, "Retrieval and feedback models for blog feed search," in *Proc. of the 31st ACM Annl. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2008, pp. 347 - 354.
[4] J Seo, WB Croft, "Blog site search using resource selection," in *Proc. of the 17th ACM Conf. on Information and knowledge management*, 2008, pp. 1053-1062.
[5] Y. Lee, S.-H. Na, J. Kim, S.-H. Nam, H.-Y. Jung, J.-H. Lee, "KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval," in *Proc. of TREC-2008*, 2009.
[6] J. Zhu, J. Wang, I.J. Cox, M.J. Taylor, "Risky business: modeling and exploiting uncertainty in information retrieval," in *Proc. of the 32nd ACM Annl. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2009, pp. 99 - 106.
[7] C. Zhai, J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. on Inf. Syst.,* vol. 22, no. 2, pp. 179 - 214, April, 2004.
[8] K. Järvelin, J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. on Inf. Syst.,* vol. 20, no. 4, pp. 422 - 446, Oct, 2002.