세 가지 정보 검색 모델의 성능 평가 및 분석
(Performance Evaluation and Analysis of Recent Information Retrieval Models)

김 지 승 † 이 준 호 †† 이 상 호 ††
(Ji Seoung Kim) (Joon Ho Lee) (Sang Ho Lee)

요 약 사용자 질의와 문서 사이의 유사도를 효과적으로 계산하는 정보 검색 모델에 대한 많은 연구가 수행되어 왔다. 이러한 연구들에 의해 개발된 정보 검색 모델들은 우수한 검색 효과를 제공한다고 알려져 있다. 그러나 이들에 대한 분석 및 검색 효과에 대한 비교 평가가 수행되지 않았기 때문에, 정보 검색 시스템의 개발시 어떠한 정보 검색 모델을 사용할 것인가에 대한 결정이 매우 어려운 실정이다. 본 연구에서는 최신 정보 검색 모델인 피벗 문서 길이 정규화, 추론 네트워크 모델, 2-포아송 모델을 분석하고, 실험을 통하여 검색 효과에 대한 비교 평가를 수행한다.

Abstract Many information retrieval models have been developed to perform effective calculation of query-document similarities. Even though it has been known that the retrieval models provide high retrieval effectiveness, little effort has been made to analyse the retrieval models and comparing their effectiveness. Hence, it is not easy to decide a retrieval model needed to develop an effective information retrieval system. In this research, we analyse the behavioral aspects of recent information retrieval models such as the pivoted document length normalization, the inference network-based retrieval model, and the 2-Possion model, and also evaluate their retrieval effectiveness through experiments.

1. 서 론

지난 40년 동안 과학과 기술 분야의 급속한 발전은 수많은 주제들에 대하여 방대한 양의 정보가 생성되는 정보화 사회를 탄생시켰다. 원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공의 여부를 결정짓는 중요한 요소가 되었다. 정보 검색 시스템은 질의에 적합한 문서들을 검색하여 사용자에게 제공함으로써, 대용량의 데이타로부터 주어진 시간 내에 원하는 정보를 발견할 수 있도록 도와준다[1].

정보 검색 시스템의 중요한 역할 중의 하나는 검색된 각각의 문서에 대하여 순위 결정 방법(Ranking)을 적용하는 것이다. 문서 순위 결정 방법은 문서와 질의 사이의 관련 정도를 나타내는 유사도(Similarity)를 계산하고, 계산된 유사도에 따라 문서에 순위를 부여한다. 높은 순위를 갖는 문서일수록 질의에 대한 만족도가 크며, 사용자는 높은 순위의 문서를 우선적으로 검토함으로써 필요한 정보를 얻는데 소모되는 시간을 최소화할 수 있다[2].

정보 검색 모델은 순위 결정 방법의 이론적 근거를 제공한다. 지금까지 사용자 질의와 문서 사이의 유사도를 효과적으로 계산하는 정보 검색 모델에 대한 많은 연구가 수행되어 왔다. 이러한 연구들에 의해 개발된 정보 검색 모델들은 우수한 검색 효과를 제공한다고 알려져 있다. 그러나, 이러한 정보 검색 모델들에 대한 분석 및 검색 효과에 대한 비교 평가가 수행되지 않았기 때문에, 정보 검색 시스템의 개발시 어떠한 정보 검색 모델을 사용할 것인가에 대한 결정이 매우 어려운 실정이다.

본 연구에서는 최신 정보 검색 모델인 피벗 문서 길이 정규화[3], 추론 네트워크 모델[4] 그리고 2-포아송 모델[5]을 분석을 수행한다. 그리고, 이들 최신 정보 검

† 비 회 원 : 숭실대학교 컴퓨터학부
jskim@searchsolutions.co.kr
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
joonho@computing.soongsil.ac.kr
shlee@computing.soongsil.ac.kr
논문접수 : 1999년 7월 6일
심사완료 : 2001년 3월 28일

색 모델들을 SMART 시스템[6]에 구현하고, 다양한 TREC 서브 컬렉션 WSJ.D2, AP.D2, ZIFF. D2, FR.D2[7]를 사용하여 이들이 제공하는 검색 효과를 비교 평가한다. 또한 이들 최신 정보 검색 모델들에 대한 분석 결과와 검색 효과 평가를 기초로 하여, 보다 높은 검색 효과를 제공하기 위하여 정보 검색 모델들이 지녀야 할 기본적인 특성들을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서 높은 검색 효과를 제공하는 것으로 알려진 피벗 문서 길이 정규화, 추론 네트워크 모델, 2-포아송 모델과 같은 최신 정보 검색 모델들 에 대하여 설명하고, 3장에서 이들에 대한 성능 평가 수행의 결과를 기술한다. 4장에서는 최신 정보 검색 모델들의 분석을 통하여, 이들이 서로 다른 검색 성능을 제공하는 이유를 설명한다. 마지막으로, 5장에서 결론을 맺는다.

2. 최신 정보 검색 모델

2.1 피벗 문서 길이 정규화

피벗 문서 길이 정규화(Pivoted Document Length Normalization)는 코사인 정규화의 문제점을 해결하기 위해 개발된 기법이다[3]. 코사인 정규화의 문제점은 다음과 같이 설명될 수 있다. 문서 집합 TREC D1 & D2와 질의 집합 TREC Q151-Q200을 대상으로 코사인 정규화를 사용하는 가중치 기법 *lnc* · *ltc*[8]를 적용하여 검색을 수행한다. *lnc* · *ltc*기법은 문서 용어 가중치 산정에는 *lnc*를 사용하고 질의 용어 가중치 산정에는 *ltc*를 사용한다는 의미로서 이를 공식으로 표현하면 다음과 같다.

• *lnc*(*document*)

$$w_{di}=\frac{\log(tf_{di})+1.0}{\sqrt{\sum_{i=1}^{n}\left[\log(tf_{di})+1.0\right]^2}}$$

• *ltc*(*query*)

$$w_{qi}=\frac{\left(\log(tf_{qi})+1.0\right)\cdot\log\left(\frac{N}{df_i}\right)}{\sqrt{\sum_{i=1}^{n}\left[\left(\log(tf_{qi})+1.0\right)\cdot\log\left(\frac{N}{df_i}\right)\right]^2}}$$

한편, TREC D1 & D2에 포함된 모든 문서들을 문서 길이의 오름차순으로 정렬한 후, 문서 1000개 단위로 구간을 나누어 각 구간을 *Bin*이라 한다. 즉, Bin_1은 가장 짧은 길이의 문서 1000개를 포함하고, Bin_i는 "$(i-1)\times1000+1$"번째 길이의 문서부터 "$i\times1000$"번째 길이의 문서를 포함한다. 이때, 질의에 대하여 적합한 문서들이 구간 Bin_i에 속할 적합 확률(Probability of Relevance)과 검색된 문서들이 구간 Bin_i에 속할 검색 확률(Pro- bability of Retrieval)은 다음과 같이 산출될 수 있다.

$$Probability\ of\ Relevance=\frac{\#\ of\ relevance\ documents\ in\ Bin_i}{\#\ of\ relevance\ documents\ in\ the\ collection}$$

$$Probability\ of\ Retrieval=\frac{\#\ of\ retrieved\ documents\ in\ Bin_i}{\#\ of\ retrieved\ documents\ in\ the\ collection}$$

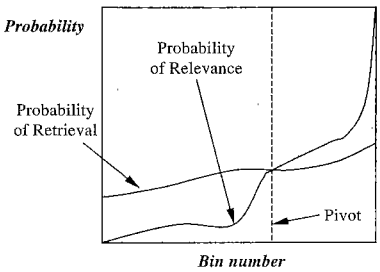


그림 1 코사인 정규화 사용시 적합 및 검색 확률의 분포

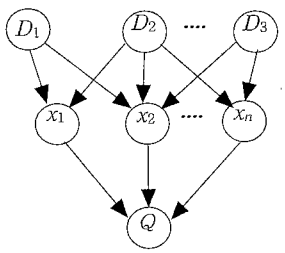


그림 2 추론 네트워크 모델

[그림 1]은 각각의 Bin_i에 대하여 계산된 적합 확률과 검색 확률의 분포 곡선을 보여준다. 이때 두 분포 곡선이 교차하는 지점을 피벗(Pivot)이라 하면, 피벗을 기준으로 문서 길이가 짧은 문서일수록 검색 확률이 적합 확률보다 높고, 문서 길이가 긴 문서들은 검색 확률이 적합 확률보다 낮음을 알 수 있다. 즉, 코사인 정규화를 사용하는 가중치 기법을 적용하여 검색을 수행할 경우, 짧은 길이의 문서가 검색시 선호된다는 것을 알 수 있다. 따라서 검색 효과를 향상시키기 위해서는 검색 확률과 적합 확률의 분포 곡선을 일치시키는 것이 바람직하며, 이를 위하여 다음과 같은 피벗 문서 길이 정규화를 사용하는 가중치 기법이 개발되었다. 다음에서 w_{di}는 문서 d에서 색인어 t_i의 가중치이며, tf_{di}는 문서 d에서 색인어 t_i의 출현 빈도이며, *slope*는 파라미터로서 그 값이 0.25일 때 가장 높은 검색 효과를 제공한다.

$$w_{di}=\frac{\dfrac{1+\log(tf_{di})}{1+\log(average\ \ tf_{di}\ in\ document\ collection)}}{(1.0-slope)+slope\times\dfrac{\#\ of\ unique\ \ terms\ in\ the\ document}{average\ \#\ of\ unique\ \ terms\ in\ document\ collection}}$$

2.2 추론 네트워크 모델

추론 네트워크 모델(Inference Network-based Retrieval Model)은 문서와 질의를 DAG (Direct Acyclic Graph)를 이용하여 표현하고, 각각의 문서가 질의를 만족하는 정도를 확률적으로 추론하는 방법이다[4]. 즉, [그림 2]에서처럼 DAG의 노드는 문서 $D_k(k=1,...,n)$, 질의 Q 그리고 문서와 질의를 표현하는 개념 $x_k(k=1...m)$이며, 링크는 노드들 사이의 관계를 나타낸다. 이때 각각의 문서 D_k가 질의 Q를 만족할 확률 $P(Q=1|D_k)$는 다음과 같이 추정될 수 있다.

$$P(Q=1|D_k)=\sum_{x_1,\ldots,x_m} P(Q=1|x_1,\ldots,x_m)\prod_{1\le i\le m} P(x_i|D_k)$$

위 식을 계산하기 위해서는 $P(Q=1|x_1,...,x_m)$ 값의 추정이 요구된다. $P(Q=1|x_1,...,x_m)$ 값의 추정 방법에 따라 새로운 연산자를 정의할 수 있으며, 이 값의 추정치는 링크 매트릭스를 사용하여 표현될 수 있다. 예를 들어, $m=3$이라 하자. 질의 Q에서 개념 x_1, x_2, x_3의 가중치가 w_1, w_2, w_3이라 할 때, 문서 D_k에서 개념 $x_i(i=1,...,3)$가 출현하는 횟수로 정의되는 weighted_sum 연산자의 링크 매트릭스는 다음과 같이 표현된다.

	$x_1x_2x_3$							
	000	001	010	011	100	101	110	111
$Q=1$	0	$\frac{w_3}{t}$	$\frac{w_2}{t}$	$\frac{w_2+w_3}{t}$	$\frac{w_1}{t}$	$\frac{w_1+w_3}{t}$	$\frac{w_1+w_2}{t}$	1

위 링크 매트릭스에서 $t=w_1+w_2+w_3$이며, 각 열은 문서 D_k에서 개념 x_1, x_2, x_3의 존재유무에 따른 $P(Q=1|x_1,...,x_m)$의 추정치를 나타낸다. 예를 들어, 문서 D_k에 개념 x_2, x_3만이 존재할 때 질의를 만족할 확률, 즉 확률 $P(Q=1|x_1=0,x_2=1,x_3=1)$의 추정치는 $(w_2+w_3)/(w_1+w_2+w_3)$이다. 위와 같은 링크 매트릭스를 사용하고, 문서 D_k에서 개념 x_i의 가중치 $P(x_i|D_k)$의 값을 p_i라 할 때, 문서와 질의간의 유사도 $P(Q=1|D_k)$다음과 같이 산출될 수 있다.

$$\begin{aligned}P(Q=1|D_k)&=\frac{w_2}{t}(1-p_1)(1-p_2)p_3+\frac{w_2}{t}(1-p_1)p_2(1-p_3)\\&+\frac{w_2+w_3}{t}(1-p_1)p_2p_3+\frac{w_1}{t}p_1(1-p_2)(1-p_3)\\&+\frac{w_1+w_3}{t}p_1(1-p_2)p_3+\frac{w_1+w_2}{t}p_1p_2(1-p_3)+p_1p_2p_3\\&=\frac{w_1p_1+w_2p_2+w_3p_3}{t}\end{aligned}$$

한편, 문서 D에서 색인어 t_i의 가중치 p_i와 질의 Q에서 색인어 t_i의 가중치 w_i는 다음과 같은 식으로 추정된다[9]. 다음에서 N은 전체 문서 수, df_i는 색인어 t_i를 포함하는 문서 수이고, tf_{di}와 tf_{qi}는 각각 문서 D와 질의 Q에서 색인어 t_i의 출현빈도이며, $maxtf_d$는 문서 D에서 최대 tf_{di}이고, H는 파라미터로서 그 값이 1.0일 때 가장 높은 검색 효과를 제공한다.

$$p_i=\left[0.4\times H+0.6\times\frac{\log(tf_{di}+0.5)}{\log(maxtf_d+1.0)}\right]\times\frac{\log\frac{N}{df_i}}{\log N}$$

$$w_i=tf_{qi}$$

2.3 2-포아송 모델

2-포아송 모델(2-Poisson Model)은 확률 검색 모델[10,11]과 포아송 분포 함수[12]를 이론적 기반으로 하여 개발된 정보 검색 모델이다[5]. 모든 문서에 대하여 용어 t_i의 출현 확률이 동일하고 임의의 문서에서 용어 t_i의 출현 빈도가 다른 문서에서 용어 t_i의 출현 빈도에 영향을 미치지 않는다면, 포아송 분포 함수를 이용하여 문서 집합에서 임의의 문서에 용어 t_i가 n번 출현할 확률을 추정할 수 있다. 문서를 구성하는 용어들은 크게 그 문서의 내용을 표현하는 주제어와 문서의 내용과 관련이 적은 비주제어로 구분될 수 있다. 비주제어는 문서 집합의 모든 문서에서 출현 확률이 동일하므로, 이러한 비주제어의 분포는 포아송 분포 함수를 사용하여 설명될 수 있다.

한편, 문서 집합에서 주제어는 특정 문서들에 집중적으로 출현하므로, 이러한 주제어의 분포는 포아송 분포를 사용하여 설명될 수 없다. 그러나, 특정 주제어가 출현하는 문서들을 그 주제어의 내용에 부합하는 문서들로 구성된 적합 클래스와 그 주제어의 내용에 부합하지 않는 문서들로 구성된 부적합 클래스로 구분하면, 각 클래스에서 주제어 출현 확률은 클래스를 구성하는 모든 문서에서 동일하며, 각 클래스에서 주제어의 분포는 포아송 분포 함수를 사용하여 설명될 수 있다. 따라서 전체 문서 집합에서 주제어의 분포는 적합 클래스에서 주제어 분포를 설명하는 포아송 분포 함수와 부적합 클래스에서 주제어 분포를 설명하는 포아송 분포 함수를 결합한 2-포아송 분포 함수로 표현될 수 있다.

2-포아송 모델은 문서 집합에서 주제어 분포를 설명하는 2-포아송 분포 함수에 확률 검색 모델을 결합하여 개발되었으며, 이 모델에서 문서 용어 가중치 공식은 다음과 같다[12]. 다음에서 N은 전체 문서 수, df_i는 색인어 t_i를 포함하는 문서 수이고, tf_{di}는 문서 d에서 색인어 t_i의 출현빈도이며, k_1과 b는 파라미터로서 각각 그 값이 2.0과 0.75일 때 가장 높은 검색 효과를 제공한다.

표 1 최신 정보 검색 모델의 유사도 및 가중치 공식

	유사도	문서 용어 가중치(w_{di})	질의 용어 가중치(w_{qi})
피벗 문서 길이 정규화	$\sum_{t=1}^{n}(w_{di}\times w_{qi})$	$\frac{\frac{1+\log tf_{di}}{1+\log(avgerage\ tf_d)}}{(1.0-slope)+slope\times\frac{\#\ of\ unique\ terms\ in\ the\ document}{average\ \#\ of\ unique\ terms\ in\ the\ document}}\times\log\frac{N+1}{df_i}$ where $slope$=0.25	$\log tf_{qi}$+1
추론 네트워크 모델	$\sum_{t=1}^{n}(w_{di}\times w_{qi})$	$\left[0.4\times H+0.6\times\frac{\log(tf_{di}+0.5)}{\log(maxtf_d+1.0)}\right]\times\frac{\log\frac{N}{df_i}}{\log N}$ where H=1.0	tf_{qi}
2-포아송 모델	$\sum_{t=1}^{n}(w_{di}\times w_{qi})$	$\frac{tf_{di}}{k_1\left((1-b)+b\cdot\frac{document\ length}{average\ document\ length}\right)+tf_{di}}\times\log\frac{N-df_i+0.5}{df_i+0.5}$ where k_1=2, b=0.75	tf_{qi}

표 2 최신 정보 검색 모델의 성능 평가 결과

	WSJ.D2		AP.D2		ZIFF.D2		FR.D2	
유사도공식, w_{di} \ w_{qi}	tf_{qi}	$\log tf_{qi}$+1	tf_{qi}	$\log tf_{qi}$+1	tf_{qi}	$\log tf_{qi}$+1	tf_{qi}	$\log tf_{qi}$+1
피벗 문서 길이 정규화	0.3455	0.3362	0.4073	0.3981	0.2052	0.2078	0.1324	0.1011
추론 네트워크 모델	0.3097	0.2563	0.3836	0.3420	0.1178	0.0569	0.1105	0.1012
2-포아송 모델	**0.3512**	0.3358	**0.4180**	0.3998	**0.2373**	0.2238	**0.1536**	0.1283

$$w_{di}=\frac{tf_{di}}{k_1\left((1-b)+b\cdot\frac{document\ length}{average\ document\ length}\right)+tf_{di}}\times log\frac{N-df_i+0.5}{df_i+0.5}$$

3. 최신 정보 검색 모델의 성능 평가

정보 검색 시스템의 검색 효과는 일반적으로 재현율(Recall)과 정확률(Precision)로써 평가된다. 재현율은 문서 집합에서 질의에 적합한 문서를 어느 정도 검색하였는가를 나타내고, 정확률은 검색된 문서들 중에서 질의에 적합한 문서가 얼마나 포함되어 있는가를 나타낸다.

본 논문에서는 재현율이 각각 0.0, 0.1, 0.2, ...,1.0일 때 정확률을 산출하고, 산출된 11개 정확률의 평균값으로 검색 효과를 측정하는 11-포인트 평균 정확률 기법을 이용하였다[6]. [표 1]에서 최신 정보 검색 모델들의 유사도 공식과 문서 용어 가중치 공식, 그리고 질의 용어 가중치 공식을 정리하였다. 이들이 검색 효과에 미치는 영향을 분석하기 위하여, 3개의 문서 용어 가중치 공식과 2개의 질의 용어 가중치 공식을 조합한 6개의 유사도 공식을 실험하였다.

즉, 6개의 유사도 공식을 구현하였으며, 테스트 컬렉션의 특수성으로 인한 부정확한 분석을 방지하기 위해 다양한 TREC 서브 컬렉션 WSJ.D2, AP.D2, ZIFF.D2, FR.D2와 100개의 TREC 질의 $Q51$-$Q150$을 사용하여 성능 평가를 수행하였다. 성능 평가를 수행한 결과, [표 2]에서 제시된 바와 같이 2-포아송 모델이 모든 TREC 서브 컬렉션에서 검색 효과가 가장 높음을 알 수 있었다.

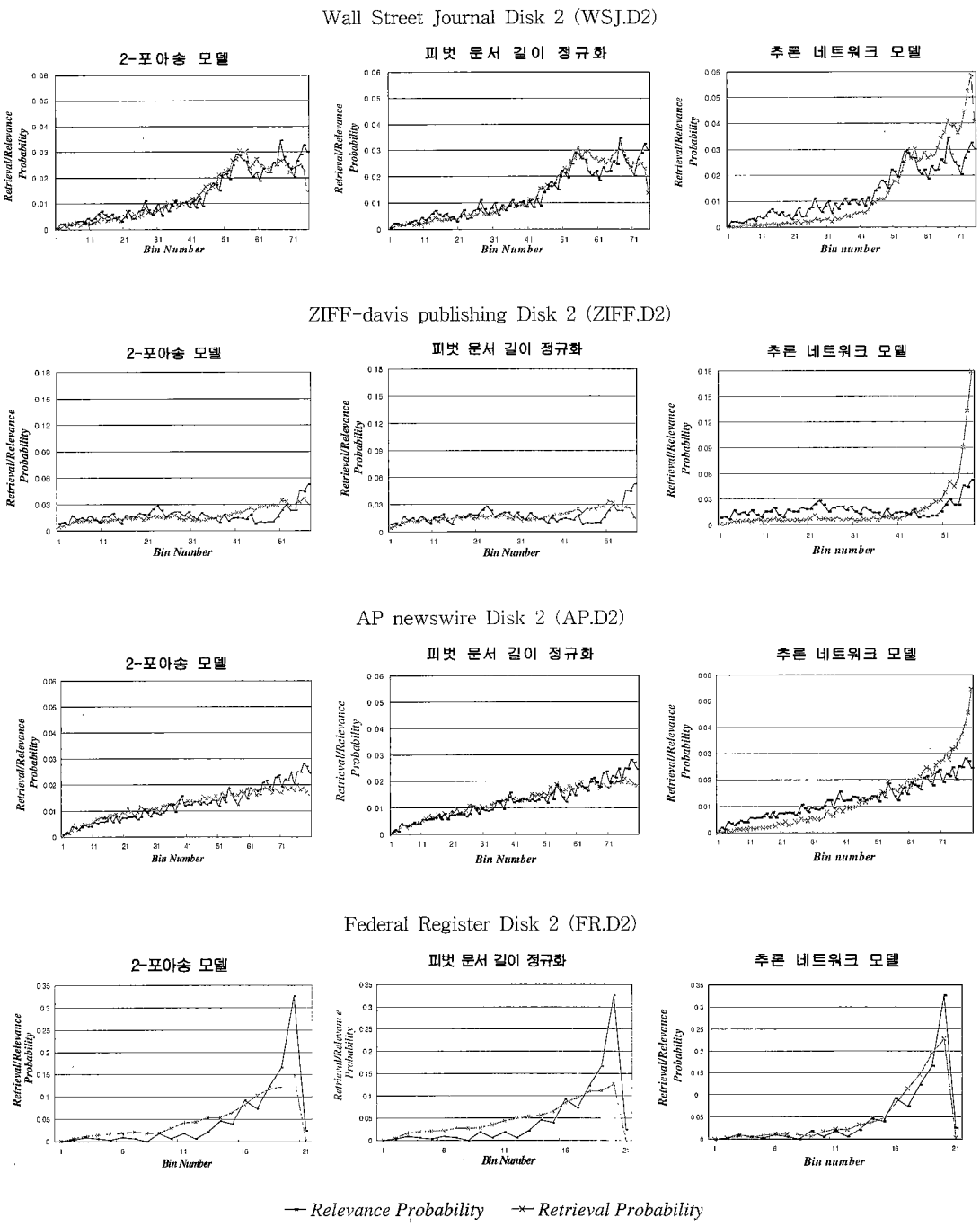


그림 3 TREC 서브 컬렉션에서 적합 및 검색 확률의 분포

4. 최신 정보 검색 모델의 분석

피벗 문서 길이 정규화 기법에서는 정보 검색 모델의 검색 효과를 향상시키기 위해서는 검색 확률 분포와 적합 확률 분포를 일치시키는 것이 바람직하다고 설명하고 있다. 즉, 검색 확률 분포가 적합 확률 분포에 근접할수록 높은 검색 효과를 제공하며, 두 분포의 격차가 커질수록 낮은 검색 효과를 제공한다는 것이다. 본 논문에서는 이러한 분석 방법을 도입하여 최신 정보 검색 모델들의 적합 확률과 검색 확률을 계산하여 [그림 3]과 같이 분포도를 작성하였다.

그 결과 WSJ.D2, ZIFF.D2, AP.D2에서는 가장 높은 검색 효과를 제공하는 2-포아송 모델의 검색 확률 분포가 적합 확률 분포와 가장 유사한 형태를 보이며, 가장 낮은 검색 효과를 제공하는 추론 네트워크 모델의 검색 확률 분포는 적합 확률 분포와 상이한 형태를 보임을 알 수 있다. 반면, FR.D2에서는 추론 네트워크 모델의 검색 확률 분포가 적합 확률 분포와 가장 유사한 형태를 보이며, 2-포아송 모델의 검색 확률 분포는 적합 확률 분포와 많은 차이가 있음을 알 수 있다.

이러한 분석을 통하여 피벗 문서 길이 정규화 기법에서 제시한 분석 방법에는 다음과 같은 문제점이 있음을 알 수 있다. 검색 확률이 적합 확률 분포에 근접할수록 높은 검색 효과를 제공할 가능성은 높아지지만, FR.D2에서와 같이 적합 확률과 검색 확률 분포 곡선이 비록 상이한 형태를 가진다 할지라도 높은 검색 효과를 제공할 수도 있다. 즉, 추론 네트워크 모델을 사용하여 검색을 수행할 경우, [그림 4]-(a)와 같이 검색된 문서들의 분포는 적합 문서들의 분포와 유사함에도 불구하고 검색된 문서들 중에서 적합 문서의 개수는 적을 수도 있다. 반면, 2-포아송 모델을 사용하여 검색을 수행하였을 경우, [그림 4]-(b)와 같이 검색된 문서들의 분포는 적합 문서들의 분포 형태와 매우 상이함에도 불구하고 검색된 문서들 중에서 적합 문서가 많이 포함되어 있을 수도 있다.

따라서, 검색 효과를 향상시키기 위해서는 검색 확률과 적합 확률 분포를 일치시키는 것이 바람직하다는 주장은 설득력을 가지지 못하며, 검색 효과를 향상시키는 다른 요인이 있음을 알 수 있다. 본 논문에서는 검색 효과를 향상시키는 다른 요인을 발견하기 위하여 최신 정보 검색 모델들의 문서 용어 가중치 공식에 대한 분석을 수행하였다. 최신 정보 검색 모델의 문서 용어 가중치 공식은 [표 1]에서와 같이 (정규화된 출현 빈도 공식×역문헌 빈도 공식)으로 구성되어 있는데, 이 중에서

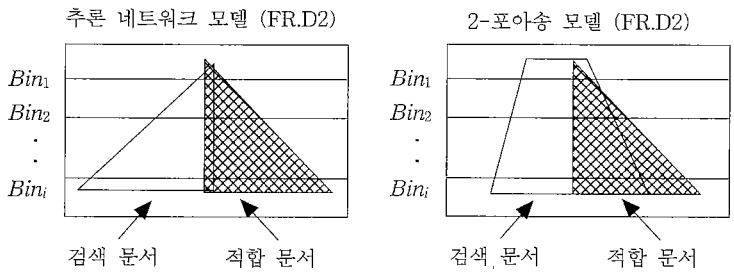


그림 4 적합 및 검색 확률 분포와 검색 효과의 관계

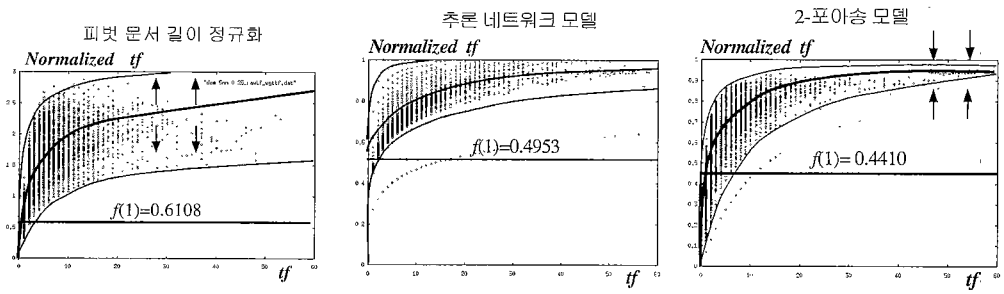


그림 5 정규화된 출현 빈도의 분포

정규화된 출현 빈도 공식을 사용하여 계산된 값으로 분포도를 작성하면 [그림 5]와 같다. 즉, 출현 빈도 *tf*의 변화에 따른 정규화된 출현 빈도(*normalized term frequency*)값을 계산하여 분포도를 작성하였다. 단, 대상 컬렉션은 WSJ.D2를 사용하였다.

4.1 피벗 문서 길이 정규화

피벗 정규화의 정규화된 출현 빈도 함수를 다음과 같이 전개해보면 2-포아송 모델의 정규화된 출현 빈도 공식과 매우 유사한 형태임을 알 수 있다.

$$s(tf)=\frac{\frac{1+\log tf}{1+\log(avgeragetf)}}{(1.0-slope)+slope\times\frac{\#\text{of uniquetermsin thedocument}}{\text{average}\#\text{of uniquetermsin thedocument}}}$$

$$=\frac{1+\log tf}{\left((1.0-slope)+slope\times\frac{\#\text{of uniquetermsin thedocument}}{\text{average}\#\text{of uniquetermsin thedocument}}\right)(1+\log(avgeragetf))}$$

$$=\frac{1+\log tf}{\left((1.0-slope)+slope\times\frac{\#\text{of uniquetermsin thedocument}}{\text{average}\#\text{of uniquetermsin thedocument}}\right)+\log(\text{avgerage}\,tf)\left((1.0-slope)+slope\times\frac{\#\text{of uniquetermsin thedocument}}{\text{average}\#\text{of uniquetermsin thedocument}}\right)}$$

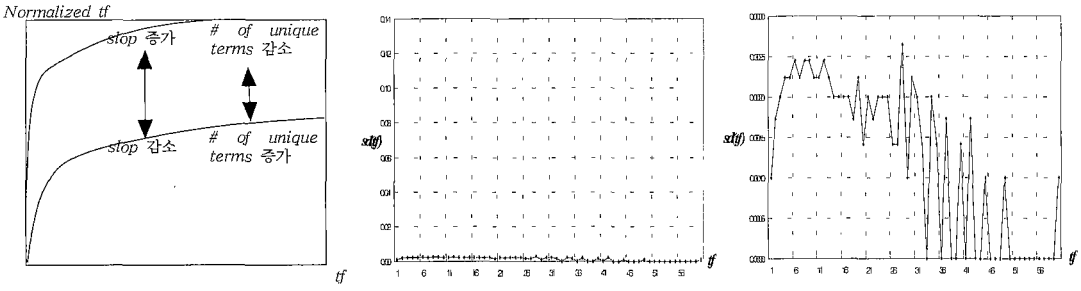
한편, [그림 5]에서와 같이 피벗 문서 길이 정규화를 사용할 경우, 출현 빈도가 0일때 정규화된 출현 빈도도 0이며 출현 빈도가 1일때 정규화된 출현 빈도는 약 0.61까지 증가한다. 정규화된 출현 빈도의 전체 범위가 0-2.31이므로, 출현 빈도가 1일때 정규화된 출현 빈도는 전체 범위의 약 26%까지 증가한다는 것을 알 수 있다. 그리고 출현 빈도가 2이상으로 증가할수록 정규화된 출현 빈도값은 완만하게 증가하는데, 이는 함수 $s(tf)$에서 $1+\log(tf)$가 출현 빈도에 대해 로그 함수를 사용하기 때문이다. 또한 $1+\log(tf)$에 로그 함수가 있으므로 tf 가 ∞로 증가할수록 함수 $s(tf)$값도 지속적으로 증가하는 특성을 가지며, 이를 식으로 표현하면 다음과 같다.

$$\lim_{tf\to\infty} s(tf)=\infty$$

또한 [그림 5]에서 출현 빈도가 동일하더라도 정규화된 출현 빈도는 상이하게 계산될 수 있음을 알 수 있다. 일반적으로 문서 길이는 # of unique terms in the document에 비례하며, # of unique terms in the document는 함수 $s(tf)$의 분모에 비례한다. 즉, 길이가 짧은 문서일 경우에는 # of unique terms in the document가 감소하므로 함수 $s(tf)$의 분모도 감소하며, 길이가 긴 문서일 경우에는 # of unique terms in the document가 증가하므로 함수 $s(tf)$의 분모도 증가한다. 그러므로 [그림 6]-(a)에서와 같이 # of unique terms in the document가 증가할수록 전반적인 정규화된 출현 빈도가 감소한다. 따라서, 동일한 출현 빈도를 갖는 용어라도 길이가 짧은 문서에 출현할 경우에는 정규화된 출현 빈도가 높게 할당되며, 길이가 긴 문서에 출현할 경우에는 정규화된 출현 빈도가 낮게 할당된다.

*slope*은 [그림 6]-(a)에서와 같이 분포의 폭을 조절

tf	1	...	6	...	12	...	∞
sd(tf)		↗		→		↘	



(a) 분포 (b) 분포의 표준 편차

그림 6 피벗 정규화 기법에서 정규화된 출현 빈도의 특성

하는 인자로서, *slope*이 증가하면 분포의 폭이 증가하며 *slope*이 감소하면 분포의 폭은 감소한다. 이때 분포의 폭이 넓어지면 길이가 짧은 문서를 선호하는 경향이 높아지며, 분포의 폭이 감소하면 길이가 긴 문서를 선호하는 경향이 높아진다. 왜냐하면, 분포의 폭이 넓어질수록 길이가 짧은 문서에 나타나는 용어에 상대적으로 높은 정규화된 출현 빈도값을 부여하므로 길이가 짧은 문서를 선호하는 경향이 나타난다.

이러한 분포 폭의 변화는 표준 편차를 계산하여 그 경향을 알수 있다. [그림 6]-(b)에서는 정규화된 출현 빈도값들의 표준 편차 $sd(tf)$를 보여주고 있는데, 좌측의 그래프에서는 다른 정보 검색 모델들과의 비교를 위해 $sd(tf)$값의 범위를 0-0.14로 설정하였다. 그 결과 표준 편차의 변화는 거의 없음을 알 수 있는데, 이를 확대하여 살펴보면 [그림 6]-(b)의 우측 그래프와 같다. [그림 6]-(b)의 우측 그래프에서 출현 빈도가 0-6일때는 표준 편차가 증가하며, 출현 빈도가 6-12일때는 표준 편차가 일정한 값을 유지한다. 그리고 출현 빈도가 13이상이 되면 표준 편차는 매우 불규칙하게 감소한다.

4.2 추론 네트워크 모델

추론 네트워크 모델의 정규화된 출현 빈도 함수를 $i(tf)$라 가정하면, 함수 $i(tf)$는 다음과 같은 식으로 표현할 수 있다. 단, H=1.0이다

$$i(tf) = 0.4 \times H + 0.6 \times \frac{\log(tf + 0.5)}{\log(max\ tf + 1.0)}$$

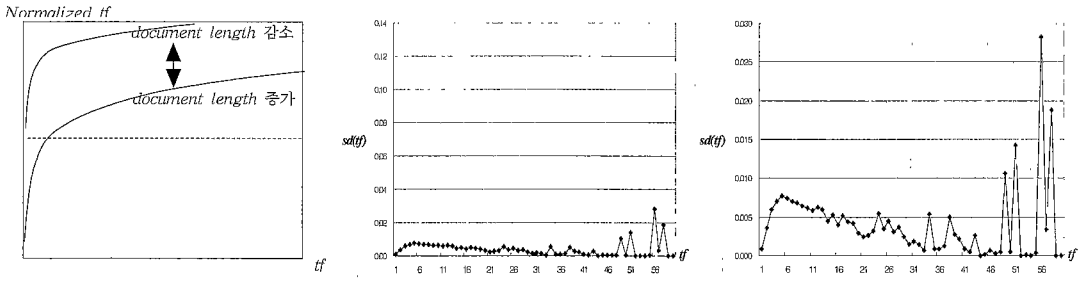
[그림 5]에서와 같이 추론 네트워크 모델을 사용할 경우, 출현 빈도가 0일때 정규화된 출현 빈도는 0이며, 출현 빈도가 1일 때 정규화된 출현 빈도는 약 0.49까지 증가한다. 정규화된 출현 빈도 값의 범위가 0부터 1정도까지이므로, 출현 빈도가 1일 때 정규화된 출현 빈도는 전체 범위의 약 49%까지 증가한다는 것을 알 수 있다. 그리고 출현 빈도가 2이상으로 증가할수록 정규화된 출현 빈도값은 완만하게 증가하는데 이러한 특성은 함수 $i(tf)$의 $\log(tf+0.5)$에서 로그 함수를 사용하기 때문이다. 또한 tf가 ∞로 증가할수록 함수 $i(tf)$값은 1로 수렴하며 이를 식으로 표현하면 다음과 같다.

$$\lim_{tf \to \infty} i(tf) = 1$$

한편, [그림 5]에서 추론 네트워크 모델의 경우에도, 출현 빈도가 동일하더라도 계산된 정규화된 출현 빈도 값은 차이가 있음을 알 수 있다. 일반적으로 문서 길이는 $max\ tf$에 비례하며, $max\ tf$가 증가할수록 함수 $i(tf)$의 분모 $\log(max\ tf+1.0)$이 증가하는 경향이 나타난다. 즉, 길이가 짧은 문서일수록 $max\ tf$가 감소하며 함수 $i(tf)$의 분모 $\log(max\ tf+1.0)$도 감소하므로 정규화된 출현 빈도값은 높아지며, 문서 길이가 긴 문서일수록 전반적인 정규화된 출현 빈도이 작아진다는 것을 알 수 있다. [그림 7]-(a)에서는 문서 길이에 따른 정규화된 출현 빈도의 변화를 그래프로 보여주고 있다.

[그림 7]-(b)에서는 정규화된 출현 빈도의 표준 편차 $sd(tf)$를 보여주고 있는데, 좌측의 그래프에서는 다른 정보 검색 모델들과의 비교를 위해 $sd(tf)$값의 범위를 0-0.14로 설정하였다. 그 결과 표준 편차의 변화는 약간 있음을 알 수 있는데, 이를 확대하여 살펴보면 [그림 7]-(b)의 우측 그래프와 같다. [그림 7]-(b)의 우측 그래프에서는 출현 빈도가 0에서 7까지는 표준 편차가 급격하게 증가하며, 출현 빈도가 7이상이 되면 표준 편차는

tf	1	...	7	...	∞
sd(tf)		↗		↘↗	



(a) 분포 (b) 분포의 표준 편차

그림 7 추론 네트워크 모델에서 정규화된 출현 빈도의 특성

불규칙하게 증가한다.

4.3 2-포아송 모델

2-포아송 모델의 정규화된 출현 빈도 함수를 $o(tf)$라 가정하면, 함수 $o(tf)$는 다음과 같은 식으로 표현할 수 있다. 단, k_1=1.0이며 b=0.75이다.

$$o(tf) = \frac{tf}{k_1\left((1-b)+b\cdot\frac{document\ \ length}{average\ \ document\ \ length}\right)+tf}$$

[그림 5]에서와 같이 2-포아송 모델을 사용할 경우, 출현 빈도가 0일 때 정규화된 출현 빈도는 0이며, 출현 빈도가 1일 때 정규화된 출현 빈도값은 평균 0.4410까지 증가한다.

정규화된 출현 빈도값의 범위가 0-1까지 이므로, 출현 빈도가 1일 때 정규화된 출현 빈도는 전체 범위의 약 44%까지 증가한다는 것을 알 수 있다. 그리고 출현 빈도가 2이상으로 증가할수록 정규화된 출현 빈도값은 완만하게 증가하는데 이러한 특성은 함수 $o(tf)$의 분모에서 tf를 사용하기 때문이다. 즉, 출현 빈도가 증가할수록 분자와 분모에 있는 tf가 증가하면서 $k_1\cdot((1-b)+b\cdot dl/avgdl)$의 비중은 작아지기 때문이다. 그러므로 출현 빈도가 증가할수록 함수 $o(tf)$값은 1로 수렴하며 이를 식으로 표현하면 다음과 같다.

$$\lim_{tf\to\infty} o(tf) = 1$$

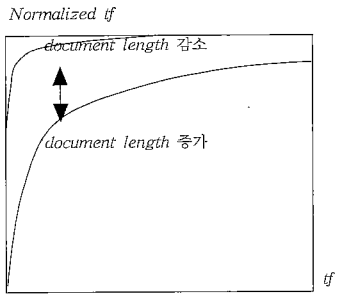


그림 8 2-포아송 모델에서 정규화된 출현 빈도의 분포

한편, [그림 5]에서 살펴본 바와 같이 2-포아송 모델의 경우에도, 출현 빈도가 동일하더라도 계산된 정규화된 출현 빈도는 차이가 있음을 알 수 있다. 일반적으로 길이가 짧은 문서일수록 함수 $o(tf)$의 $k_1\cdot((1-b)+b\cdot dl/avgdl)$은 감소하므로 정규화된 출현 빈도는 증가하며, 길이가 긴 문서일수록 $k_1\cdot((1-b)+b\cdot dl/avgdl)$은 증가하므로 정규화된 출현 빈도는 감소한다. 따라서, [그림 8]에서와 같이 문서 길이가 증가할수록 정규화된 출현 빈도는 전반적으로 낮아진다는 것을 알 수 있다.

인자 b는 0에서 1사이의 값을 갖는 상수로서, 분포의 폭을 조절하는 역할을 한다. 즉, b값이 감소할수록 분포의 폭은 좁아지므로 길이가 긴 문서를 선호하는 경향이 높아지며, b값이 증가할수록 분포의 폭은 넓어지므로 길이가 긴 문서를 선호하는 경향이 제거된다. 예를 들어 다음과 같이 문서 d_1, d_2, 그리고 질의 q_1이 있다고 가정하자.

$$d_1 = \{(t_1, 1), (t_2, 1), ..., (t_n, 1)\}$$
$$d_2 = \{(t_1, 2), (t_2, 2), ..., (t_n, 2)\}$$
$$q_1 = \{(t_1, w_1), (t_2, w_2), ..., (t_n, w_n)\}$$

이때 문서 d_2는 d_1을 두 번 중첩되어 표현된 것이므로 질의 q_1에 대해 동일한 유사도를 생성하는 것이 바람직하다. 다음에서는 b=0.1과 b=1.0인 경우, 문서와 질의 사이의 유사도 변화를 살펴본다. 단, k_1값은 1.0으로 설정하였다.

b=0.1을 적용하면 가중치 공식은 다음과 같이 변환된다.

$$o(tf) = \frac{tf}{\left(0.9+0.1\cdot\frac{document\,length}{1.5n}\right)+tf}$$

위의 가중치 공식을 이용하여 문서 d_1, d_2과 질의 q_1과의 유사도를 계산하면 다음과 같다.

$$Sim(d_1, q_1) = 0.51\cdot w_1 + 0.51\cdot w_2 + ... + 0.51\cdot w_n = 0.51\cdot\sum_{i=1}^{n} w_i$$
$$Sim(d_2, q_1) = 0.66\cdot w_1 + 0.66\cdot w_2 + ... + 0.66\cdot w_n = 0.66\cdot\sum_{i=1}^{n} w_i$$

따라서, b=0.1을 적용할 경우에는 길이가 긴 문서가 선호됨을 알 수 있다.

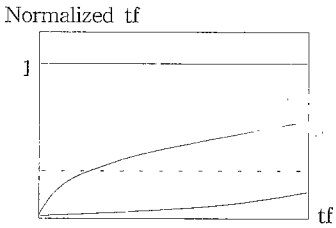
한편 b=1.0을 적용하면 가중치 공식은 다음과 같이 변환된다.

$$o(tf) = \frac{tf}{\frac{document\ \ length}{1.5n}+tf}$$

위의 가중치 공식을 이용하여 문서 d_1, d_2과 질의 q_1과의 유사도를 계산하면 다음과 같다.

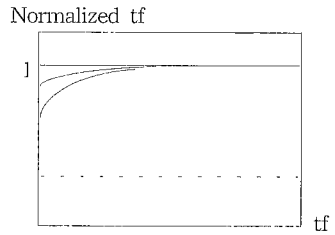
$$Sim(d_1, q_1) = 0.6\cdot w_1 + 0.6\cdot w_2 + ... + 0.6\cdot w_n = 0.6\cdot\sum_{i=1}^{n} w_i$$
$$Sim(d_2, q_1) = 0.6\cdot w_1 + 0.6\cdot w_2 + ... + 0.6\cdot w_n = 0.6\cdot\sum_{i=1}^{n} w_i$$

따라서, b=1.0을 적용할 경우에는 두 문서가 동일하게 취급됨을 알 수 있다.



k_2이 증가할수록

그림 9 인자 k_1의 특성



k_2이 감소할수록

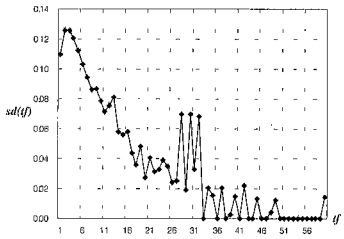


그림 10 2-포아송 모델에서 정규화된 출현 빈도의 표준 편차

인자 b가 용어의 출현 빈도에 관계없이 분포의 전반적인 표준 편차를 조절하는 역할을 하는 것과는 달리, 인자 k_1은 용어의 출현 빈도에 따라 상이한 표준 편차를 가지도록 조절해주는 역할을 한다. 낮은 출현 빈도를 갖는 용어를 t_{low}라 하고 높은 출현 빈도를 갖는 용어를 t_{high}라 하면, [그림 9]에서와 같이 인자 k_1이 감소할수록 t_{low}의 표준 편차는 t_{high}의 표준 편차보다 상대적으로 커지는 반면, 인자 k_1이 증가할수록 t_{low}의 표준 편차는 t_{high}의 표준 편차보다 작아진다. 이와 같이 2-포아송 모델의 가중치 공식은 정규화된 출현 빈도의 분포 특성을 정교하게 조절할 수 있는 검색 모델이다.

한편, [그림 10]에서는 2-포아송 모델에서 정규화된 출현 빈도의 표준 편차 $sd(tf)$의 분포를 보여주고 있는데, 다른 정보 검색 모델보다 표준 편차의 변화 정도가 매우 크다는 것을 알 수 있다. 출현 빈도가 0에서 2까지는 표준 편차가 0.1258정도까지 증가하며, 출현 빈도가 2이상이 되면 표준 편차가 급격하게 감소함을 알 수 있다. 즉, 용어 t_i의 출현 빈도가 1, 2, 3 정도일 경우에는 용어 t_i가 길이가 긴 문서에 출현하는지 또는 길이가 짧은 문서에 출현하는지에 따라 정규화된 출현 빈도값을 다르게 할당하고, 출현 빈도가 증가할수록 문서 길이에 관계없이 용어 t_i에 정규화된 출현 빈도값을 할당한다.

tf	1	....	2	....	∞
$sd(tf)$		↗		↘	

5. 정규화된 출현 빈도 함수의 특성

4장에서 최신 정보 검색 모델의 특성을 분석한 결과, 검색 효과 개선을 위하여 정규화된 출현 빈도 함수 $f(tf)$로는 다음에 제시한 조건을 만족해야 한다.

5.1 함수 f(tf)는 f(0)=0, f(1)∈[0.3, 0.5] 조건을 만족해야 한다.

테스트 컬렉션상의 문서에서 용어 출현 빈도가 1인 경우가 전체 용어의 약 70-90%를 차지하며, 출현 빈도가 2인 경우가 전체 용어의 약 5-10%를 차지하고, 출현 빈도가 3-60인 용어의 비중은 매우 작게 나타난다. 즉, 출현 빈도값의 범위는 1-60까지의 넓은 분포를 가지지만 테스트 컬렉션상에서 용어들의 출현 빈도값은 대부분 1 또는 2이므로, 출현 빈도의 높고 낮음만으로는 문서의 적합성 여부를 판단하기 어렵다.

그러므로 출현 빈도의 높고 낮음보다는 용어의 출현 여부가 문서의 적합성 여부를 판단하기 위한 중요한 요소라는 것을 알 수 있다. 이러한 특성을 최신 정보 검색 모델에서도 반영하고 있으며 [그림 5]에서는 이를 분포로서 나타내고 있다. 최신 정보 검색 모델의 정규화된 출현 빈도값은 출현 빈도가 1일 때 매우 급격하게 증가하며, 그 증가량은 정규화된 출현 빈도값의 전체 범위에서 약 30-50%정도를 차지한다.

최신 정보 검색 모델은 문서의 검색 여부를 결정하기 위해 용어의 출현 빈도보다, 오히려 출현 유무에 높은 비중을 둔다는 것을 알 수 있다. 여러 개의 질의 용어를 이용하여 검색을 수행한다고 가정할 때, 특정 질의 용어가 많이 출현하는 문서보다는 여러 개의 질의 용어들을 포함하고 있는 문서를 적합한 문서로 간주한다. 즉, 정

보 검색 모델이 높은 검색 효과를 제공하기 위해서는 출현 빈도가 1일때 정규화된 출현 빈도값이 급격하게 증가하는 특성을 가지는 문서 용어 가중치 공식을 사용해야 한다.

5.2 함수 f(tf)는 로그 형태의 분포 특성을 가져야 한다.

최신 정보 검색 모델의 정규화된 출현 빈도값은 출현 빈도가 작을 때는 급격하게 증가하고, 출현 빈도가 높아질수록 정규화된 출현 빈도값이 완만하게 증가하는 로그 형태의 분포를 가진다. 즉, 출현 빈도가 낮은 용어들 간에는 출현 빈도의 많고 적음이 문서의 검색 여부에 영향을 미칠수 있지만, 출현 빈도가 높아질수록 출현 빈도의 많고 적음은 문서의 검색 여부에 영향을 미치지 않는다는 것을 나타낸다.

문서에 어떤 용어가 출현하지 않은 경우보다 1번 출현한 경우에 높은 가중치를 문서에 부여해야 하는 것은 분명하다. 그러나 용어 t의 출현 빈도가 문서 d_1에서는 50, 문서 d_2에서는 1일 경우에, 문서 d_1보다 문서 d_2에 50배의 가중치를 부여하는 것은 부적절하다. 왜냐하면 출현 빈도가 매우 높은 용어는 극소수에 불과하며, 이러한 용어는 오히려 불용어일 가능성이 높기 때문이다. 보다 높은 검색 효과를 제공하기 위해서는 정규화된 출현 빈도값의 분포는 로그 형태의 분포 특성을 가지는 것이 바람직하다.

5.3 함수 f(tf)는 tf→∞일때 상수 C로 수렴해야 한다.

$$\lim_{tf\to\infty} f(tf) = C$$

최신 정보 검색 모델의 정규화된 출현 빈도값의 분포는 상수값으로 수렴하는 특성을 갖는다. 출현 빈도가 일정 횟수를 초과하면 높은 출현 빈도로서의 중요도가 제거되는 경향을 가지므로, 정규화된 출현 빈도가 증가할수록 일정한 값을 유지하는 것이 바람직하다.

5.4 함수 f(tf)의 표준 편차는 tf→∞일때 0으로 수렴하며 다음과 같은 분포 특성을 가져야 한다.

$$\lim_{tf\to\infty} sd(tf) = 0$$

tf	1	....	2	....	∞
$sd(tf)$		↗		↘	

표준 편차의 변화 특성은 높은 검색 효과를 제공하기 위하여 정보 검색 모델이 지녀야 할 기본적인 특성들 중에서 가장 핵심적인 요소이다. 정규화된 출현 빈도값의 분포에서 표준 편차는 동일한 출현 빈도값을 가지는 용어일지라도, 문서 길이에 따라 서로 다른 정규화된 출현 빈도값을 부여하는 역할을 한다. 즉, 동일한 출현 빈도값을 가지는 용어일지라도 긴 문서에서 출현한 용어에는 낮은 정규화된 출현 빈도값을 부여하고, 짧은 문서에서 출현한 용어에는 높은 정규화된 출현 빈도값을 부여한다.

예를 들어, 각각 100개와 20개의 용어를 포함하고 있는 긴 문서 d_1과 짧은 문서 d_2가 있고, 용어 t가 문서 d_1과 d_2에서 각각 10번씩 출현한다고 가정하자. 이때 용어 t는 d_2에서 절반의 비중을 가지는 중요한 용어로서 취급되는 것이 바람직하지만, 문서 d_1에서 용어 t의 비중은 10%에 불과하다. 그러므로 동일한 출현 빈도값을 가지는 용어일지라도 그 용어가 출현하는 문서의 길이에 따라 서로 다른 정규화된 출현 빈도값을 부여하는 것이 바람직하다.

정규화된 출현 빈도값의 분포에서 출현 빈도가 낮을 때는 표준 편차가 매우 크며, 출현 빈도가 높아질수록 표준 편차는 감소한다는 것을 알 수 있다. 추론 네트워크 모델의 표준 편차는 출현 빈도가 증가할수록 증가하는 경향을 보이며 성능 평가 결과 가장 낮은 검색 효과를 제공하므로, 출현 빈도가 높아질수록 표준 편차가 감소하는 특성이 검색 효과를 개선시키는 요소로서 작용한다는 것을 알 수 있다.

왜냐하면 출현 빈도가 낮은 용어일수록 문서의 적합성 판정에 큰 영향을 주기 때문에 문서 길이가 중요한 요소로 작용한다. 그러나 출현 빈도가 높은 용어일수록 문서의 적합성 판정에 영향을 미치지 않으므로, 용어가 출현한 문서의 길이는 검색 수행시에 고려하지 않는 것이 효과적이다. 검색 효과를 개선하기 위해서는 문서 길이에 따라 서로 상이한 가중치를 부여해야 하며, 출현 빈도가 높아질수록 문서 길이는 고려하지 않는 것이 바람직하다.

최신 정보 검색 모델들 중에서 가장 높은 검색 효과를 제공하는 2-포아송 모델의 분포 특성이 앞서 제시한 분포 특성에 가장 부합된다. 특히 출현 빈도가 매우 작은 경우에는 표준 편차가 매우 크며, 출현 빈도가 높아지면서 표준 편차도 낮아지는 특성은 2-포아송 모델의 검색 효과를 개선시키는데 큰 역할을 한다. 그러나 피벗 문서 길이 정규화 기법이나 추론 네트워크 모델의 표준 편차 분포는 상대적으로 매우 낮은 수준에 불과함을 알 수 있다. 그리고 추론 네트워크 모델은 첫 번째 조건만을 만족하고 나머지 특성들은 만족하지 않았으므로 가장 낮은 검색 효과를 제공한다. 본 연구에서 제시한 정규화된 출현 빈도값의 분포 특성들은 검색 효과 개선을 위한 필수적인 요소이다. 요컨대, 정보 검색 모델의 문서 용어 가중치 공식에 상관없이, 그 가중치 공식이 본 연구에서 제시한 분포 특성을 가지면 높은 검색 효과를 제공할 수 있다.

6. 결 론

보다 효과적인 정보 검색 시스템의 개발을 위하여, 사용자 질의와 문서 사이의 유사도 계산을 효과적으로 수행하는 정보 검색 모델에 대한 많은 연구가 수행되어 왔다. 이러한 연구들에 의해 개발된 최신 정보 검색 모델인 피벗 문서 길이 정규화, 추론 네트워크 모델, 2-포아송 모델은 우수한 검색 효과를 제공한다고 알려져 있다. 그러나, 이들 최신 정보 검색 모델들에 대한 분석과 검색 효과에 대한 비교 평가가 수행되지 않았기 때문에, 정보 검색 시스템의 개발시 어떠한 정보 검색 모델을 사용할 것인가에 대한 결정이 매우 어려운 실정이다.

본 연구에서는 이들 최신 정보 검색 모델들에 대한 분석을 수행하고, 실험을 통하여 검색 효과에 대한 비교 평가를 수행하였다. 최신 정보 검색 모델에 대한 성능 평가를 수행한 결과, 테스트 컬렉션에 관계없이 2-포아송 모델이 가장 우수한 검색 효과를 제공함을 알 수 있었다. 또한 문서 용어 가중치 공식에서 정규화된 출현 빈도가 검색 효과에 가장 큰 영향을 미치는 요소임을 알 수 있었다. 따라서 정규화된 출현 빈도에 대하여 중점적으로 분석을 수행하였다. 이를 통하여 보다 높은 검색 효과를 제공하기 위하여 정보 검색 모델들이 지녀야 할 기본적인 특성들을 추출함으로써 정보 검색 모델의 선정 및 개발에 대한 기준을 제시하였다. 향후 연구 과제로 문서 용어 가중치 기법의 역문서 빈도와 질의 용어 가중치 기법에 대한 구체적인 분석이 남아있다. 또한 보다 높은 검색 효과를 제공하기 위하여 정보 검색 모델이 지녀야 할 기본적인 특성을 만족하는 정보 검색 모델 개발이 필요하다.

참 고 문 헌

[1] Salton, G., "Historical note: The past thirty years in information retrieval," Journal of the American Society for Information Science, Vol. 38, No. 5, pp. 375-380, 1987.

[2] Lee, J.H., Kim, M.H. and Lee, Y.J., "Ranking documents in thesaurus-based Boolean retrieval systems," Information Processing & Management, Vol. 30, No. 1, pp. 79-91, 1994.

[3] Singhal, A., Buckley, C. and Mitra, M., "Pivoted document length normalization," Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21-29, 1996.

[4] Turtle, H. and Croft, W.B., "Evaluation of an inference network-based retrieval model," ACM Transactions on Information Systems, Vol. 9, No. 3, pp. 187-222, 1991.

[5] Robertson, S.E. and Walker, S., "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232- 241, 1994.

[6] Salton, G. and McGill, M.J., Introduction to Modern Information Retrieval, McGraw-Hill, Inc., 1983.

[7] Harman, D., "Overview of the 1st text retrieval conference," Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 36-48, 1993.

[8] J.H. Lee, "Combining multiple evidence from different properties of weighting schemes," Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 180-188, 1995

[9] Greiff, W.R., Croft, W.B. and Turtle, H., "Computationally tractable probabilistic modeling of Boolean operators," Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 119-128, 1997.

[10] Maron, M.E. and Kuhns, J.L., "On relevance, probabilistic indexing and information retreival," Association for Computing Machinery, Vol. 7, No. 3, pp. 216-244, 1960.

[11] Robertson, S.E. and Sparck Jones, K., "Relevance weighting of search terms," Journal of the American Society for Information Science, Vol. 27, pp. 129-146, 1976.

[12] Harter, S.P., "A probabilistic approach to automatic keyword indexing," Journal of the American Society for Information Science, Vol. 26, pp. 197-206, 1975.

[13] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M. and Gatford, M., "Okapi at TREC-3," Proceedings of the 3rd Text REtrieval Conference (TREC-3), pp. 109-126, 1995.

김 지 승

1996년 숭실대학교 산업공학과(학사). 1999년 숭실대학교 컴퓨터학과(석사). 관심분야는 정보검색모델, 정보검색시스템

이 준 호

1987년 서울대학교 컴퓨터공학과(학사). 1989년 한국과학기술원 전산학과(석사). 1993년 한국과학기술원 전산학과(박사). 1993년 ~ 1994년 한국과학기술원 인공지능연구센터 연구원. 1994년 ~ 1995년 코넬대학교 전산학과 방문연구원. 1994년 ~ 1997년 연구개발정보센터, 선임연구원. 1997년 ~ 현재 숭실대학교 컴퓨터학부 조교수. 관심분야는 정보검색, 데이타베이스

이 상 호

1984년 서울대학교 컴퓨터공학과(학사). 1986년 미국 노스웨스턴 대학교 전산학과(석사). 1989년 미국 노스웨스턴 대학교 전산학과(박사). 1990년 ~ 1992년 한국전자통신연구소 데이타베이스실. 1992년 ~ 현재 숭실대학교 컴퓨터학부 부교수. 1999년 ~ 2000년 미국 George mason 대학교 방문교수. 관심분야는 인터넷 데이타베이스, 데이타베이스 성능평가, 트랜잭션 처리