베이지안 추론방, 기반 색인어의 심층 분석 방법
승서광, 이승우, 정한민
한국과학기술정보연구원, SW연구실
{esmail, swlee, jhm}@kistl.re.kr

Deep Analysis on Index Terms Using Baysian Inference Network
Sa-kwang Song, Seungwoo Lee, Hanmin Jung
Korea Institute of Science and Technology Information, Dept. of SW Research

요 약
대부분의 검색 엔진에서 색인어의 추출 및 기술적 부여방법은 매우 중요한 연구주제로, 검색 엔진의 성능은 주요 영향을 미친다. 일반적으로, 검색어 리스트를 통해 성능에 긍정적인 영향을 미치지 않는 색인어를 제거하려니, 핵심어가 아닌 전문용어 등 상대적으로 중요한 색인어를 강조하는 방식을 사용하여 검색엔진의 성능을 향상시킨다. 하지만, 이론 분야, 형태소 분석, 검색어 처리 등 검색엔진의 단계별 처리 과정에서, 개별적인 색인어가 검색엔진에 미치는 영향을 분석하고 이를 반영한 검색 엔진 성능 향상 기법론 제시되지 않고 있다. 따라서 본 연구에서는 각 단계별 처리 과정에서 생성된 색인어에 미치는 영향을 계량화하여 긍정적/부정적 색인어를 분류하는 방법론을 소개하고, 이를 기반으로 색인어 기술의 조합함으로써 검색 엔진의 성능 향상 가능성을 제시한다.

주제어: 오류 분석, 색인어 가중치, 추론방, 정보검색

1. 서론


이러한 검색 엔진 연구 및 개발 과정의 어려움을 감소시키고 생산성 향상을 지원하기 위해 본 논문에서는 검색 엔진의 세부 모델별 분석 결과를 자동으로 가시화하고 각 세부 처리 결과의 개별 개체가 성능에 미치는 영향을 계량하여 제시하였다. 이를 통해, 검색 엔진 세부 모듈들의 업출력에 따른 성능을 구체적으로 살펴볼 수 있는 방법을 제공한다.

논문의 구성은 다음과 같다. 2장에서는 검색 엔진의 오류 분석에 대한 관련 연구를 소개하고 3장에서 본 논문에서 제안하는 추론방 기반 색인어 가중치 분석 방법을 설명하고, 4절에 결론을 끝낸다.

2. 관련연구

대부분의 검색 엔진 연구는 검색 시스템의 다양한 부소 모듈을 분석 및 수정함으로써 성능향상을 평가하고 해도 과연이 아니다. 이때 주로 표준 테스트 컬렉션이나 과거의 관련 실험 결과를 활용하여 오류 분석을 진행하게 되는데, 연구자 또는 개발자의 각자 경험에 의하여 ad-hoc한 방식으로 진행되는 것이 일반적이다. 따라서, 다양한 오류를 분석하고 이를 성능향상에 반영하기 위한 일반화된 방법론이나 분석 시도가 쉽지 않은 게 현실이 다.

하지만 미국의 NIST에서 6주 동안 진행한 RIA(Reliable Information Access)라는 워크샵[8]에서 매우 중요한 시도가 있었다. 복수의 검색 엔진의 결과를 분석하여 시스템의 다양성을 요인과 주제 요인을 분석하고 이들이 검색엔진 성능에 어떻게 기여하는지를 정리하였다. TREC과 같은 컬렉션을 활용한 실험에서 검색의 주제 및 시스템 종류에 따라 성능 차이가 매우 두드러진 현상을 발견할 수 있었다. 이는 기존적인 검색 엔진의 경우뿐만 아니라, 검색 확장(query expansion)과 같은 기능을 포함한 진보된 검색 엔진에서도 동일하게 나타나는 것을 확인할 수 있었다. 참고로, RIA 워크샵에
서는 7개의 성능이 검증된 최고의 검색 엔진 및 그 그룹이 참여하여 실험을 진행하였다. 또한 28명의 전문가 및 12개 기관으로부터 다수의 대학원생들이 참여하였고 검색 엔진의 오류 분석은 주로 수직계에 의존하여 진행하였다. 다만 각 그룹별 분석 지원 도구를 활용하였는데, Waterloo 대학의 User Interface(WUI), NST의 Beadplot, Clairvoyance의 AWB, SMART시스템의 SMART_retro 등이 그 예이다.

WUI는 한 검색 시스템의 검색 결과 문서를 분석하는 데 사용된 주요 도구로서, 사용자가 선택한 점의 토대로 하여 적합/부적합 여부 또는 검색되었는지 여부 등을 제공하는 단순한 기능을 제공한다. 이러한 단순한 기능을 제공함에도 많이 쓰이게 된 이유는 적합/부적합, 검색 여부 등이 검색 결과를 판단하고 분석하는데 가장 기본적인 정보이고, 이를 단순한 인터페이스로 제공하기 때문이다.

다음의 Beadplot은 두 개 이상의 검색 결과를 한눈에 볼 수 있도록, 색상 패턴을 활용하여 비슷한 검색 결과를 제공하는 검색시스템이 유사하게 보이도록 가시화하는 도구이다. 그림 1은 Beadplot을 통해 8개의 검색 결과를 가시화한 그림으로, 각 행은 각 검색 엔진의 검색 결과 문서에 대응한다. 기준 행으로 첫 번째 검색 결과와 접합의 정렬된 문서들이 색상 스크립트를 따라 제시되고, 각 문서가 다른 검색 결과에서 어느 위치에 랭크되는지는 적절적으로 볼 수 있도록 가시화한 것이다. 그림 1에서 각 행에 분포되어 있는 같은 모양의 박스들은 같은 문서를 의미하고, 해당 문서의 위치는 그 문서의 검색 행렬원측부터 순서대로 나타낸다.

다음으로, AWB는 파이프라인으로 구성된 임의의 템플로 문서를 그룹화하는 기능을 제공한다. 예를 들어, 상위에 랭크된 검색 문서의 80%에서 나타나고, 동시에 검색되지 않은 문서의 5%에서 출현하는 섹터를 제시하는 기능을 제공한다.

마지막으로 SMART_retro는 특별한 기능을 제공하는 데, 특정 질의에 의해서 검색된 모든 문서 점검에서 최적의 범위를 구현하고 주는 기능을 제공한다. 이 최적의 범위를 이용하여 검색을 수행한 결과 MAP(Mean Average Precision) 0.83로 매우 높은 성능을 얻을 수 있었다.

이러한 도구들은 나름대로 검색엔진의 오류를 분석할 때 유용하게 사용할 수 있지만, 단편적이고 전체적인 정보를 제공하는 선에 다물고 있다. 따라서, RIA 워크샵에서는 이러한 도구들은 검색 시스템 및 주제별 분석을 위한 작업 수행 시에 제한적으로 밖에 사용되지 않았고 많은 부분은 개별 연구자나 개발자간의 경험이나 복잡한 예외를 의지한 분석이 이루어졌다.

따라서, 본 연구에서는 연구자 또는 개발자들에게 검색 엔진 내부의 개별 분석 모듈들에 대한 좀 더 구체화된 정보를 제공하고 이를 통한 검색 성능 향상도 지원할 수 있는 방안으로, 검색 모델을 많이 활용하는 추론 맵[3,5,6,9]을 활용한 검색 엔진 오류 분석 지원 방안에 대해 소개한다.

3. 본론
3.1 추론망 기반 검색 엔진 내부 분석

그림 2 다중 레이어로 구성된 페이지안 추론 검색엔진의 세부 처리 단계 모식도

검색 엔진은 보통 일련의 단위 처리들의 모음으로 구성되어 있다. 각 단위 처리는 입력된 질의의 속성을 이용하여 결과 문서의 집합을 추정하고, 그 중에서도 최적의 문서 집합을 추정한다. 이러한 단위 처리들은 검색 엔진의 최근의 동향을 반영하고, 서로 다른 방법의 결과를 비교하여 보다 정확한 검색 결과를 제공한다. 

성된다. 예를 들어, 문장 분리, 어절 분리, 불용어 처리, 스테밍(Stemming) 등의 일련의 작업을 통해 대상 문서를 분석하고 색인어를 추출하는데 때 단계 입력과 출력 개체가 정해지고, 입력과 출력을 연결하면 그림 2와 같은 입력의 내트워크가 형성된다. 그림 2에서 문서 처리의 각 단계(행)는 노드들의 집합으로 구성되고 이전 단계의 출력노드에서 다음 단계의 입력노드로의 연결이 이루어지는데, 이 외의 문서로부터 이와 같은 내트워크를 구성하더라도 이 연결의 효과를 부여할 수 있게 된다. 이것은 메시지란 추론망 모델에서 사용하는 Belief 확률값(3)이 된다.

예를 들어, 그림 3에서 '문장1'에서 '어절1'로의 연결, '어절2'에서 '색인어1'과 '색인어2'로의 연결 이에 해당된다. 이 Belief 확률값은 상위 노드에 연결되어 있는 하위 노드들 중에서 특정 노드에 연결될 확률을 식 1과 같이 계산한 것으로, 일반적으로 문서 레이어와 색인어 레이어 두개로만 구성되어 본 연구에서는 각 세부 모듈들의 미출력 관계를 명확히 하고자 레이어를 확장 가능하게 구성하였다.

\[ Belief(N_i) = db + (1 - db) TF_i \cdot IDF_i \]  
(식 1)

여기서 \( N_i \)는 \( k \)-번째 노드, TF는 노드의 Term Frequency, IDF는 노드의 Inverse Document Frequency를 의미한다. 이러한 방식으로 구현되어진 시스템이 그림 3인데, 이 예는 어절분리, 불용어처리, 스테밍 (Stemming) 레이어로 구성되었고, 마지막에 진의어결과 연결되어 있다.

일단 모든 노드간의 Belief 값이 구해지면 이를 이용해 내트워크가 구성되고, 특정 점집이 주어지게 되면, 어떤 문서와 해당 점집이 어떻게 연결되어 있는지를 확인할 수 있게 된다. 즉, 주어진 점집이에 대해 각 문서와의 유사도가 어떻게 계산되고 어떤 값을 가지며 개별 점집의 어절어 집단의 문서와 연관 관계를 냅는지 등 다양한 정보가 가시적으로 표현된다.

이때 각 노드 간에 연결된 Belief값은 의도적으로 수정 또는 삭제가 가능하며 특정 두 노드 간에 연결된 Belief값을 변경하면, 결과적으로 점집의 대비 문서의 유사도가 변경되어, 해당 문서의 검색 성능이 변경되는 결과를 가져온다. 이러한 기능은 검색 모델 연구가 또는 개발자에게 매우 적절적이면서도 세분화된 분석을 가능하게 하는 기능이다. 예를 들어, 스테밍(Stemming) 레이어에서 특정 노드 추출 방식을 변경하였을 경우, 이 변경 방식이 전체 시스템 성능에 어떤 영향을 미치는지 파악하려면 활용하거나, 특정 패턴의 노드 간의 Belief 값을 일괄적으로 변경함에 따른 성능 비교 등이 가능해진다.

3.2 노드의 분별력

각 노드는 적합한 문서에서 추출될 수도 부합하도록 문서에서 추출될 수도 있기 때문에 적합 문서와 부합함 문서 집합에서의 각 노드의 Belief값의 차이를 이용해 해당 노드 문서를 적합부합으로 구분하는 역할을 정의하였다. 그림 4와 같이, 이는 특정노드의 Belief 값이 적합 문서 집합에서와 부합함 문서 집합에서 두드리지 차이가 난다면, 해당 노드는 문서들의 적합성을 결정할 때 매우 유용한 노드로서 다른 노드와 구분될 필요가 있다. 이 가정을 확신히 하는 가정은 Song(2012)[10]은 거의 가정을 기반으로 PD(Prediction Power)을 설계하고, 이를 복수의 검색 모델에 적용하여 검색 성능을 높이는 연구를 수행하였다.

이 차이를 해당 노드의 분별력이라고 정의하고 이 분별력을 기본으로 노드들간의 가중치가 일정 값에 지정되지 않다. 이 경우 적합하고 부합하여 사용자에게 제공한다. 이 분별력은 해당 노드가 세부 모듈들을 관리하게 주어진 것인지 그렇지 않은지에 대한 간접적인 정보를 제공하게 된다. 이 분별력이 작은 노드는 세부 노드 내의 아이의 중요도의 적용을 위한 단계로 활용하여 결론적으로 전체 시스템 성능에 영향을 미칠 수 있다. 이때, 노드의 분별력 계산방법은 Song(2012)[10]을 따른다. 이러한 분별력에

이 분별력을 기준으로 정정적 노드와 부정적 노드를 구분한 것이 표 1이다. 표 1의 제목에서 볼 수 있는 것처럼 일반적으로 각 레이어에서 추출된 어절 그룹이 정정적 그룹과 부정적 그룹 사이에 큰 차이를 보임을 알 수 있다.

<table>
<thead>
<tr>
<th>레이어</th>
<th>정정적 노드 (분별력&gt;0)</th>
<th>부정적 노드 (분별력&lt;0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>토론 분리</td>
<td>Amtrak, dogs, Exxon, dog, NRA, Dow, gel, breast, implant, Corning, silicone, implants</td>
<td>methods, method, Methods, methodical, Method, Methodism, methodically, METHOD, methodical, Mathematical, METHODS, Methode</td>
</tr>
<tr>
<td>불용어 처리</td>
<td>dog, exxon, cessation, mauled gel, dow, breast, nra, implant, corning, silicone, implants</td>
<td>real, instance, instances, instance, instanced, instancing, methods, method, methodical, methodism, methodically, methodic, method</td>
</tr>
<tr>
<td>스테밍 (Stemming)</td>
<td>medic, valdez, concern, exxon, devic, corn, implant, nra, breast, dow, gel, silicon</td>
<td>grant, stop, judgement, comput, electron, involv, describ, refer, accomplish, real, instanc, method</td>
</tr>
</tbody>
</table>

5. 결론

본 논문에서는 검색 엔진 인구자나 개발자들이 검색 모델 또는 검색엔진의 세부 모듈에서 발생할 수 있는 내부의 오류 가능성 확인을 도울 수 있는 방법을 제안하였다. 특히, 베이지안 추론방식 기반으로 검색엔진의 세부 모듈과 그에 따른 입출력 노드의 정보를 상세히 확인할 수 있는 GUI 기반의 분석 시스템을 제공하여 사용자가 필요에 따라, 특정 노드의 값을 수정/삭제 등의 작업할 수 있고 그에 따른 결과를 파악할 수 있는 방법을 제공하였다.

특히, 각 노드의 적합 문헌 집합과 부적합 문헌 집합에서의 가중치 차이를 기반으로 노드의 분별력을 계산하고 이를 분별력에 기반하여 색인어 간의 차별성을 보여 함으로써, 세부 노드별 중요도를 제공한다.

주의 연구로는 제한된 도구의 사용성 평가를 수행하며, 사용자들의 의견을 수렴하고 수렴한 의견을 기준으로 사용자 친화적인 분석 기능을 추가하고자 한다.

참고문헌

[3] D. Metzler, and Croft, W.B., Combining the Language Model and Inference Network Approaches to Retrieval, Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5), 735-750, 2004
[4] D. Harman, What we have learned, and not learned, from TREC. In the Proceedings of the 22nd Annual Colloquium on Information Retrieval Research, pages 2-21, 2000