수식 관계를 이용한 검색 결과 랭킹 시스템과 향상된 검색 엔진 인터페이스를 통한 검색 과정의 효율성 향상

문욱성, 최주원*
대구고등학교
한국과학기술원 전자전산학과 전산학전공*
ukseong@gmail.com, jwchoi@world.kaist.ac.kr*

Search Ranking System Using Modification Relation and Improved Search Engine Interface to Enhance Search Experience

Ukseong Moon, Joo-Won Choi*
Daegu High School
Division of Computer Science, KAIST*

요 약

본 논문에서는 현재 검색 엔진의 랭킹 방식의 문제점과 인터페이스의 문제점을 해결하기 위하여 노력하였다. 기존의 페이지간 링크와 같은 부가적 정보를 이용한 인기도 기반 랭킹의 문제점을 단어간의 수식 관계를 이용한 의미 기반 랭킹 알고리즘의 제시를 통해 해결하였다. 또한 검색어와 연관된 단어를 수식 관계를 이용하 계산, 시각화하여 제공함으로써 사용자가 잘못된 검색어로 검색을 시작하였더라도 항상 올바른 검색 결과를 얻을 수 있도록 도왔으며 각 검색 결과와 함께 원문을 요약해 제공함으로써 검색 결과를 일일이 클릭해 보지 않고도 내용을 쉽게 유추할 수 있도록 도왔다.

1. 서 론

방대한 양의 정보에서 원하는 정보를 효과적으로 검색하는 것은 정보화 시대를 살아가는 사람들에게 중요한 요소이다. 이러한 검색의 중요성에 따라 검색엔진들은 사용자가 원하는 정보를 효과적으로 전달하려 하고 있다. 하지만 현재의 검색 엔진은 페이지의 랭킹 방식과 인터페이스에서 여러 문제점을 가지고 있다.

첫째, 현재의 검색 시스템의 랭킹 방식은 문서의 의미적 정보를 이용하지 않고, 페이지의 인기도를 이용한다. 그러나 페이지의 인기도를 분석하기 위해서는 웹페이지들 간의 상호 링크 정보 등의 많은 부가적 정보를 필요로 한다. 또한 의미정보를 이용하지 않기 때문에 사용자가 검색하고자 하는 정보와는 동떨어진 결과가 상위에 랭크될 수도 있다. 따라서 웹페이지를 검색할 때 페이지의 의미 정보를 이용하면 좀 더 적절한 검색결과를 얻을 수 있을 것이다.

둘째, 현 검색 시스템은 사용자가 항상 적절한 검색어를 알고 있어야 원하는 정보를 검색 할 수 있으나, 그렇지 못한 경우가 많다. 따라서 검색 시스템이 사용자를 올바른 검색어로 유도할 수 있다면 사용자의 검색 능력이 부족하여도 사용자가 필요로 하는 정보를 제공할 수 있다. 따라서 사용자의 검색어와 연관된 정보를 제공하면 사용자를 적절한 검색어로 유도할 수 있다.

마지막으로, 현재의 검색 시스템의 검색 결과만으로는 본문의 내용을 유추할 수 없는 경우가 많아 사용자가 해당 페이지로 이동하여 페이지에서 자신이 필요한 내용인지 분석해야 한다. 따라서 웹페이지 원문의 내용을 요약하여 검색결과로 보여주면 요약 결과를 바탕으로 원래 페이지의 내용을 유추할 수 있게 되어 앞에서의 불편함을 방지할 수 있게 된다.

본 논문에서는 페이지의 의미정보를 이용하여 페이지를 랭크하여 사용자에게 필요한 내용을 포함한 페이지가 상위에 랭크될 수 있도록 하였다. 또한 연관 검색어를 올바른 검색어로 유도하였으며, 검색 결과로 각 페이지의 요약된 내용을 보여줌으로써 각 페이지의 내용을 좀 더 쉽게 유추할 수 있도록 하였다.

2. 기존의 검색 시스템 분석

2.1. 기존의 랭킹 알고리즘

기존의 랭킹 알고리즘은 문서 외적인 부가정보를 이용하여 각 페이지의 중요성을 판단한다.

이러한 랭킹 알고리즘의 잘 알려진 방법으로는 웹페이

지의 백링크[1]의 개수를 바탕으로 순위를 매기는 방법[1]과 아마존에서 제품을 추천하기 위해 사용자의 구매 횟수를 이용하는 방법[2]이 있다.

이런 방식은 각 페이지의 의미적 정보를 반영하지 못하므로 검색 결과의 적절성에 한계가 있으며 웹페이지들간의 링크 관계 등의 부가적 정보를 제공하지 못하는 환경에서는 적용할 수 없으므로 일반적인 텍스트 검색에 적용할 수 없다는 단점이 있다.

2.2. 기존의 검색 인터페이스

현재의 검색 인터페이스는 단순히 제목과 각 페이지의 일부분만 보여주므로 사용자가 원하는 결과를 얻기 위해서는 각 페이지에 들어가서 자기가 원하던 정보인지 확인하는 절차가 필요하다.

또한 어떤 단어로 검색하냐에 따라 원하는 결과를 얻을 확률이 많이 달라진다. 따라서 효율적인 검색을 위해서는 일정 수준 이상의 검색 능력을 필요로 하며 실제 실험에서도 사용자의 검색 능력에 따라 효율성의 편차도 큰 것으로 나타났다.

따라서 본 논문에서는 새로운 방식의 검색 랭킹 알고리즘과 함께 새로운 검색 엔진 인터페이스 제안을 통해 사용자의 검색 능력과 관계없이 효율적인 검색을 할 수 있도록 하였다.

3. 의미 정보에 기반한 검색 시스템

새로운 랭킹 알고리즘의 특징은 문서 그 자체의 내용에 기반하였다는 것이다. 따라서 적용 범위의 한계를 극복할 수 있으며 내용 그 자체에 기반하는 것이므로 그 적절성 또한 보장된다.

새로운 랭킹 알고리즘에서는 문서의 내용에 기반한 랭킹을 하기 위하여 문서 내 단어간의 수식 관계를 이용한다. 수식 관계를 이용하면 문장에서 중요하게 생각하는 단어가 무엇인지 알 수 있고, 이를 이용하여 사용자가 검색한 단어가 중요한 비중을 차지하는 문서를 상위에 랭크하게 되면 사용자가 원하는 내용을 담은 문서에 사용자가 쉽게 접근할 수 있게 된다.

3.1. 단어간 수식 관계에 기반한 랭킹 알고리즘

수식 관계에 기반한 랭킹을 위한 첫 번째 과정은 각 문서의 단어간 수식 관계를 분석하는 것이다. 한 단어가 수식을 많이 받았다 함은, 그 단어에 대한 부가 정보가 많이 필요했다는 것이다. 따라서 그 단어는 문서 내에서 핵심적 내용을 담고 있을 확률이 높다.

단어간 수식 관계가 분석[2]되면 이 데이터를 이용하여 각 명사, 약어, 그리고 서수들의 수식 단어 개수를 구한다. 수식 단어의 개수를 셀 때에는 직접적인 수식 외에도 간접적인 수식까지도 모두 검사하여 내용 전반에 걸쳐 중요한 단어를 찾아내게 된다. 각 단어의 수식어의 개수를 체크한 뒤에는 수식어의 개수를 전체에서의 비율로 모두 변환한다. 그 뒤 문서에서 주요 문장을 추출(3.2 참조)하여 주요 문장에서 출현하는 횟수를 비율화하여 해당 비율만큼 원래의 수식어 개수에 가중치를 준다[그림 2]. 이를 통해 전체 문서에서 동등하게 각 단어의 문서에서의 중요성을 비교할 수 있게 된다.

Algorithm to rank search results

```
IMPORTANCE_OF_WORDS( target_document, main_sentences )
01  words << all the words of a target_document
02  modify << ∮
03  possibles << ∮
04  frequency << ∮
05  importance_rates << ∮
06  rate_of_possible << 0
07  for each words:
08    if word ∈ noun, abbreviation, or cardinal numbers then
09      add word to possibles
10      modify[word] << | modifiers of word |
11      frequency[word] << | word in main_sentences |
12    end if
13  end for
14  for each possibles:
15    rate_of_possible << ( modify[possible] / sum of modify )
16    importance_rates[possible] << rate_of_possible +
                rate_of_possible( frequency[possible] / sum of frequency )
17  end for
18  return importance_rates
END IMPORTANCE_OF_WORDS
```

[그림 1] 수식 관계를 이용한 단어 별 문서 내 중요도 측정 알고리즘

수식 관계를 통해 문서에서 추가적인 정보를 많이 주려고 노력한 단어가 무엇인지 알 수 있으므로 핵심 단어를 효율적으로 판단할 수 있다. 따라서 수식어의 개수와 주요 문장에서의 출현 빈도를 이용하여 계산된 각 단어의 중요도를 이용하면 사용자가 원하는 내용을 가진 문서를 상위에 랭크 할 수 있다.

$$\text{Word Importance} = \frac{|M_W|}{\sum |M_W|} + \frac{|M_W|}{\sum |M_W|}\left(\frac{|F_W|}{\sum |F_W|}\right)$$

M_W= { x | x is modifiers of W }

F_W= { x | x is W in main sentences }

[그림 2] 수식 관계를 이용한 단어 별 문서 내 중요도 측정 수식

3.2. 단어간 수식관계를 이용한 연관성 그래프

현재의 검색 시스템의 문제점은 사용자가 잘못된 단어로 검색을 하고 있더라도 그 사용자를 올바른 검색 결과로 유도할 방법이 없다는 것이다.

본 논문에서는 이런 문제를 해결하기 위하여 각 검색 결과 페이지의 측면에 검색한 단어와 연관된 단어들을 결과 문서들에서 찾아내어 그 중 주요한 단어들간의 연관성을 시각화하여 그래프로 보여주고, 그 그래프를 사용자가 확장해 나갈 수 있도록 하였다. 이를 통해 사용자는 쉽게 현재 단어와 연관된 정보를 얻을 수 있어 잘

---

[1] 백링크(back link: inedges): 특정 페이지를 가리키는 링크
포워드링크(forward link: outedges): 특정 페이지에서 밖으로 나가는 링크

[2] 수식 관계의 분석은 Connexor Parser를 이용함

못된 단어로 검색하고 있더라도 정확한 단어를 좀 더 손쉽게 찾아 갈 수 있을 것이다.

단어간의 연관성을 계산하기 위해서는 검색된 문서에 포함된 단어 중 출현 빈도가 높은 단어를 연관된 단어로 추출한다. 이 때에 해당 단어가 사용자가 검색한 단어와 수식 관계가 존재한다면 가중치를 주어 추출될 확률을 높이게 된다.

$$Relatedness = \sum |D_W| + \sum |D_W| \left(\frac{\sum |D_{M_T}|}{\sum |D_W|} \right)$$

D_W= { x | x is documents contain W and a target word }
D_{M_T}= { y | y is documents that W is in modifiers of a target word }

[그림 3] 수식 관계를 이용한 단어의 연관도 측정 수식

Algorithm to find related words

```
RELATEDNESS_OF_WORDS( target_word )
01  words << words of documents that contain target_word
02  relatedness << 0
03  word_contains << 0
04  modifiers << ∮
05  modify << 0
06  related << ∮
07  for each words:
08    if word ∈ noun, abbreviation, or cardinal number then
09      word_contains << documents contain word
11      for each word_contains:
12        modifiers << modifiers of target_word in word_contains
13        if word is in modifiers then
14          modify << modify + 1
15        end if
16      end for
17      relatedness << | word_contains | +
18                     | word_contains | ( modify / | word_contains | )
19      if relatedness > threshold then
20        add word to related
21      end if
22      modify << 0
23    end if
24  end for
25  return related
END RELATEDNESS_OF_WORDS
```

[그림 4] 수식 관계를 이용한 연관 단어 추출 알고리즘

단어를 포함하고 있는 문서들의 단어들의 출현 빈도와 각 문서 내에서의 수식 관계를 이용하여 특정 단어의 연관된 단어를 추출할 수 있다.

3.3. 중요 문장 요약을 통한 검색 결과 출력 시스템

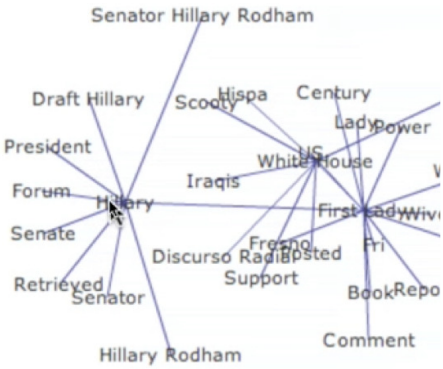
현재의 검색 엔진에서는 검색된 결과 중에서 중요한 내용을 사용자가 직접 분석하여 찾아내야 하여 사용자는 검색 단어를 생각해 내는 데 들어가는 노력 외에도 자신에게 중요한 정보를 찾고 그 정보를 재구성하는 추가적인 노력을 투자하여야 한다.

본 논문에서는 검색 결과를 분석하는 단계의 효율성 향상을 위하여 검색 결과의 중요한 내용을 분석하고, 요약하여 사용자에게 제공하였다. 이를 통해 사용자는 검색 결과를 분석하는 노력을 줄일 수 있어 원하는 정보를 얻는 과정을 효율적으로 진행할 수 있다.

요약된 내용을 제공하기 위해서는 첫 번째로 중요한 문장을 추출한다. 이를 위해 각 문장을 구문 분석하여 각 단어의 의존관계를 분석한다. 분석된 의존 관계의 루트로부터 거리가 가까울수록 해당 단어는 높은 의존도를 갖게 되며 의존도가 높은 단어들이 속한 문장들을 골라 축약하게 된다.

두 번째로는 중요 문장을 축약하는 것이다. 이를 위하여 각 문장을 구문 분석한 뒤 각 단어가 요약된 내용에 포함되는지 등을 분석해 주제와의 연관성을 계산하여 높은 연관성을 가진 단어들을 골라 문장을 축약한다.

이렇게 요약된 내용을 검색 결과에 표시하여 사용자가 각 검색 결과의 페이지로 이동하지 않아도 각 결과의 내용을 분석하여 원하는 내용이 있는 결과가 무엇인지 분석하고 더 필요한 내용을 쉽게 분별할 수 있다.



[그림 5] 단어간 연관성 그래프

Flex를 이용하여 제작된 그래프를 이용하여 단어간 연관성을 사용자가 효율적으로 사용할 수 있도록 시각화 할 수 있다. 시각화된 단어간 연관성을 이용하여 각 단어에 대한 주요 정보를 효율적으로 얻을 수 있으며 클릭을 통해 추가적으로 각 단어의 연관된 단어들에 접근할 수 있다. 이를 통해 내가 필요한 정보를 쉽게 골라낼 수 있고 따라서 사용자가 잘못된 단어로 검색을 시작하였더라도 사용자를 올바른 검색 결과로 유도할 수 있다.

4. 실험 및 평가

본 논문에서는 2와 3에서 제시한 새로운 랭킹 알고리즘과 인터페이스를 검증하기 위하여 Goocle[3]이란 툴을 제작하여 사용자가 이 툴을 이용하여 검색하는 과정을 Google을 이용하여 검색하는 과정과 비교하였다.

그러나 데이터의 부족과 인터페이스 상의 사소한 차이점 때문에 본 논문에서 제시한 랭킹 알고리즘과 인터페이스를 적용한 Goocle을 Google과 비교하는 단순 비교하는 것에는 문제점이 많았다. 우선 저자의 이전 논문에서 시행하였던 새로운 인터페이스를 적용한 툴을 이용한 실험에서 검색 시간은 약 15%, 검색의 정확도는 약 20% 정도가 사용자의 검색 능력과 상관없이 일정하게 늘어났다는 것을 토대로 이번에 시행할 실험도 기존의 검색 엔진보다 뛰어난 결과를 보여줄 수 있을 것이라

[3] http://goocle.ukseong.com/

는 것을 추정할 수 있다.

5. 결 론

본 논문에서는 단어간 수식 관계를 이용하여 새로운 방식의 의미 기반 랭킹 알고리즘을 제안하였다. 또한 수식 관계를 이용해 생성된 단어간 연관성을 그래프를 통해 제공하여 사용자를 올바른 검색어로 유도하였으며, 중요 문장 추출과 문장 요약을 이용하여 검색 결과로 요약된 정보를 제공해 사용자가 각 페이지의 내용을 쉽게 유추할 수 있도록 하였다. 실험 데이터의 수집 부족으로 현재 본 논문에서 제시한 새로운 알고리즘과 방식들을 일반 검색 사이트와 비교하는 것은 어렵지만 이전의 유사한 실험 결과를 통하여 예측되는 결과는 새로운 인터페이스와 랭킹 알고리즘이 사용자의 검색 과정의 효율성을 향상시킬 수 있다는 것이다. 또한 본 시스템은 웹뿐만이 아닌 어느 문서의 검색에도 보편적으로 적용 가능하다는 장점이 있어 일반 문서에 대해서도 의미정보를 이용한 검색을 적용할 수 있게 될 것이다.

참고 문헌

[1] Haveliwala, T. "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search", IEEE Trans. Knowl. Data Eng., 15(4):784--796, 2003.

[2] Linden, G. et al. “Amazon.com recommendations: item-to-item collaborative filtering”, Internet Computing, IEEE, Volume 7, Issue 1, Jan.-Feb. 2003 Page(s):76 - 80

감사의 말

이 논문의 집필을 위한 연구 과정을 도와주신 한국과학기술원의 최기선 교수님과 한국과학기술원 시맨틱웹 연구센터 관계자 여러분께 감사의 말씀을 드립니다.