문장 분석 및 온톨로지를 이용한 Focused Crawler

최광복[O] 김현주 강진범 홍광희 양재영 최중민
한양대학교
{kbchoi[O], hjkim, jbkang, khhong, jyyang, jmchoi}@cse.hanyang.ac.kr,

Focused Crawler using Ontology and Sentence Analysis

Kwangbok Choi[O], Hyunjoo Kim, Jinbeom Kang, Kwanghee Hong, Jaeyoung Yang, Joongmin Choi
Hanyang University

요 약

월드 와이드 웹의 보편화로 인하여 급속하게 증가하고 변화하는 웹 문서는 검색엔진으로 하여금 색인된 웹 문서와 현재의 웹 문서의 일관성을 유지할 수 없을 정도이다. 이러한 문제를 해결하기 위한 방법으로 연구되고 있는 것이 특정한 주제를 정하고 정해진 주제에 관련된 문서를 수집할 수 있는 focused crawler가 제시되고 있다. 지금까지 다양한 접근방법의 focused crawler가 개발되었지만, 모두 웹 링크를 이용하여 연결되어 있는 문서를 평가 하는 처리과정을 거치고 있다. 그러나 이러한 과정은 다양한 내용을 포함하고 있는 문서일 경우 관련내용이 존재함에도 문서가 버려지거나 사용되더라도 문서상의 모든 링크를 사용하여 처리하는 비효율적인 문제점이 발생한다. 이 논문에서는 웹 문서 내부에 포함되어 있는 정보를 온톨로지를 이용하여 평가함으로써 다양한 내용을 가진 문서에서 사용자가 원하는 정보를 찾을 수 있을 뿐만 아니라 정보와 관련된 링크만을 사용하여 보다 효율적이고 정확한 문서를 수집하고자 한다.

1. 서론

월드 와이드 웹이 보편화 되면서 웹은 사용자들이 필요한 정보를 수집할 뿐만 아니라 지식이나 정보를 저장하고 이를 공유하기 위해 사용되는 거대한 지식베이스가 되었다. 이러한 방대한 양의 정보를 빠르고 정확하게 찾기 위해서 정보검색 시스템이 중요한 역할을 수행한다. 현재 검색 서비스를 제공하고 있는 곳에서는 수백 기가 바이트에서 테라바이트의 데이터를 보유하며, 이미지 검색에서 지도검색까지 다양한 서비스를 제공하고 있다. 하지만 아직도 사용자의 질의에 의해 검색되어지는 정보는 사용자의 정보요구를 만족 시키지 못하는 경우가 발생하고 있는데, 이것은 유사한 정보가 출력되거나 사용자의 의도와는 다른 정보들이 출력되는 경우에서 발생한다.

이러한 상황에서 최근 연구가 되고 있는 것이 focused crawler이다.[1,2,5,6] 웹에서는 하루가 다르게 새로운 정보들이 새로 추가 되거나 혹은 사라지며 갱신되고 있다. 이것은 모든 웹 페이지들을 다시 수집하고 색인하는 것이 불가능해지고 있음을 의미한다. Focused crawler는 방대한 웹에서 특정 주제에 대한 문서만을 수집하게 하여 수집된 정보와 현재의 웹 문서의 일관성을 유지할 수 있도록 한다. focused crawler에도 크게 두 가지 문제점이 존재하고 있는데, 첫 번째 현재의 focused crawler는 기존의 crawler와 마찬가지로 한번은 문서상에서 존재하는 모든 하이퍼링크를 방문하여 문서를 다운받아 분석하고 관계없는 문서는 폐기한다. 이것은 관련성이 없는 링크를 방문하는 횟수를 증가시키게 되며 관련 없는 문서도 문서를 평가하는 시간이 추가적으로 필요함을 의미한다. 두 번째 문서의 평가에 있어서 문서 전체를 하나의 단일 유닛으로 생각하기 때문에 연관성이 없다고 판단되는 경우에는 문서 전체가 버려지는 경우가 발생한다. 이것은 관련성 있는 정보가 웹 문서에 존재함에도 불구하고 문서 전체의 관련성 여부에 의해서 버려진다.

이 논문에서는 이러한 문제점을 해결하기 위해 하나의 문서를 여러 단위체로 나누어 분석하고 관련성 없는 하이퍼링크만을 제거하는 방법을 제안한다.

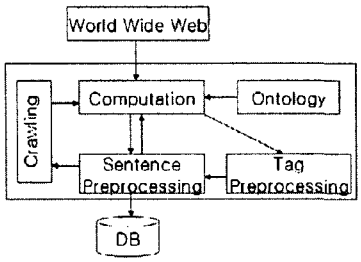
2. 웹 문서 분류

이 논문에서는 웹 문서를 크게 하이퍼링크만 존재하는 웹 문서와 절에 하이퍼링크가 존재하는 웹 문서로 구분한다. 예를 들어 야후와 같이 디렉토리가 하이퍼링크로 연결되어 있는 페이지에서는 기존의 crawler가 사용한 방식을 따르며 절이나 문장에 하이퍼링크가 존재하는 문서는 이 논문에서 제안한 분석방법을 통해서 분석하게 된다.

일반적으로 절과 절은 전체적인 의미는 같을 수 있으나 다루는 세부 주제는 다르다. 따라서 각각의 절을 문서에서 추출해내고 분석할 수 있다면 문서 전체의 의미는 사용자의 정보요구와 다르더라도 세부 주제를 나타내는 절을 이용하여 정보의 손실을 줄일 수 있다.

3. 시스템 구조

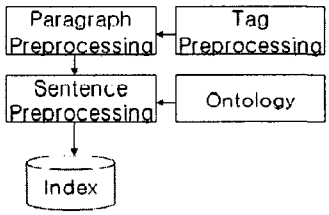
이 논문에서 제안한 focused crawler에서 사용하고 있는 구조는 기존의 문서를 수집하는 방식과는 다르게 그림 1에서 볼 수 있는 것처럼 온톨로지를 사용하여, 웹 문서가 입력되었을 때 문서 내용을 평가하게 된다.



<그림 1. Focused Crawler 구조>

4. 웹 문서 처리

문서 전체를 평가했을 때 결과 값이 기준치 이상 나올 경우 문서상에 사용자가 원하는 정보가 존재 한다고 판단 할 수 있다. 이 경우 어떠한 부분에서 사용자가 원하는 정보가 존재하고 있는지를 확인하고 필요한 정보만을 얻기 위해서는 그림 2와 같은 구조로 문장의 평가가 이루어져야 한다.



<그림 2. 웹 문서 처리과정>

4.1 문서 평가

사용될 문서에 대해 주제와의 관련성을 평가시, 다음과 같은 수식을 이용하였다.[3]

$$P = \frac{M}{Ws}$$

Ws = 온톨로지에 사용된 중복되지 않는 모든 단어

M = 매칭된 단어

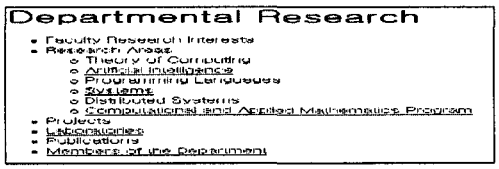
문서에서 어떠한 부분이 사용자가 원하는 정보를 가지고 있는지 알 수 없기 때문에 문서 전체를 대상으로 매칭된 단어의 수를 구하고 전체 온톨로지의 개수로 나누게 된다. 연산결과가 기준치 이상일 경우 평가된 문서 내에서 관련된 문장을 찾기 위한 과정이 수행된다.

문제점으로써 이 온톨로지에 정의 되어진 단어수가 적을 경우 관련은 없지만 유사한 단어가 사용된 문서도 포함될 가능성이 높아진다.

4.2 태그 처리

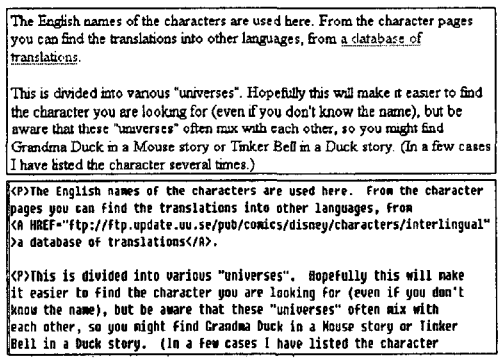
웹 문서는 기본적으로 HTML을 사용하여 문서를 작성하고 있으며, 기존 연구에서는 HTML 태그들을 모두 제거하고 있지만, 이 논문에서는 몇 가지 태그를 사용하여 문서를 처리 할 때 사용하게 된다. 예를 들면, 그림 3과 그림 4와 같이 HTML문서에서 문단을 나눌 때 "<p>"혹은 "

..."와 같은 형식이 주로 사용되고 있으며, 문서의 순서나 주제별 분류를 나타낼 때 "<li>", "<u>" 등을 사용하고 있다. 이 논문에서는 이러한 태그를 웹 문서에서 문장을 분리해내기 위해서 사용한다.




```
<li>Research Areas<ul>
  <li> <a href="http://theory.cs.uchicago.edu">
   Theory of Computing</a>
  <li> <a href="http://ai.cs.uchicago.edu">Artificial
   Intelligence</a>
   .....
  </ul> </li>
```

<그림 3. 주제를 분류하는 웹 문서와 태그>



The English names of the characters are used here. From the character pages you can find the translations into other languages, from a database of translations.

This is divided into various "universes". Hopefully this will make it easier to find the character you are looking for (even if you don't know the name), but be aware that these "universes" often mix with each other, so you might find Grandma Duck in a Mouse story or Tinker Bell in a Duck story. (In a few cases I have listed the character several times.)

```
<P>The English names of the characters are used here.  From the character
pages you can find the translations into other languages, from
<A HREF="ftp://ftp.update.uu.se/pub/comics/disney/characters/interlingual"
>a database of translations</A>.

<P>This is divided into various "universes".  Hopefully this will make
it easier to find the character you are looking for (even if you don't
know the name), but be aware that these "universes" often mix with
each other, so you might find Grandma Duck in a Mouse story or Tinker
Bell in a Duck story.  (In a few cases I have listed the character
```

<그림 4. 문단을 분류하는 웹 문서와 태그>

4.3 문단 처리

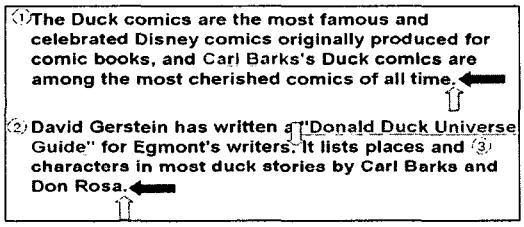
"<p>"와 같은 태그를 이용하여 각각의 문단으로 나누고, 평가대상인 문장을 찾기 위해서 문장별로 나누어야 한다. 그러기 위해서 이 논문에서는 "."를 이용하여 문장을 나누고 있다. 하지만 이러한 처리방식에도 몇 가지 문제가 있다. 이 중 가장 큰 문제는 사용자의 문서 작성의 불규칙성이다. 웹 문서의 경우 HTML은 태그 외 문서작성에 규제가 없기 때문이다. 두 번째로 웹 주소나 e-mail의 경우도 "."를 포함하고 있기 때문에 분석에 어려움이 있다.

때문에 이 논문에서는 웹 주소와 e-mail과 같이 패턴 분석이 가능한 경우에만 문단의 하위 구조인 문장처리를 하고 있다.

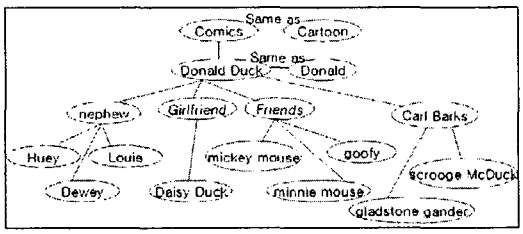
4.4 문장 처리

앞에서 언급했듯이 하나의 문서는 다양한 세부주제를 가지는 문단이나 문장으로 구성되어있다. 이와 같은 경우 사용자가 요구하는 내용을 포함하는 부분을 찾아내는 것이 가장 중요하다. 사용자의 정보요구에 관련성 있는 부분을 인식하기 위해서 이 논문은 문장 분석을 이용한다.

문단 처리를 거치게 되면 그림 5에서 보는 것과 같은 ①,②,③의 3문장으로 분류된다. 이 경우 하이퍼링크가 연결되어 있는 ①,②가 문장 분석에 사용된다. 이 두 문장을 평가 할 때 그림 6과 같은 온톨로지를 이용하여 문장의 관련성을 측정하게 된다.



<그림 5. 문장 분류예제>



<그림 6. 온톨로지>

예를 들어, ②문장을 평가한다면, 문서 전체에서 각각의 단어들을 뽑아내어 그림 6의 온톨로지를 이용하여 검증한다면 "Donald Duck"이라는 단어가 매칭 될 것이다. 하지만 단어 하나로 평가 문장을 결정지을 수 없다. 자신이 속한 온톨로지 구조에서 상위 혹은 하위의 연관된 단어도 함께 존재 해야만 한다.
그림 7은 문장을 평가 할 때, 온톨로지에서 정의한 단어를 이용하여 평가하고 관련성 높은 문장의 하이퍼링크만을 Queue에 입력하는 알고리즘을 나타내고 있다.

```
While(not End(ontology)){                    setIndex(element){
    Onto = setIndex(ontology.element())          Array list;
    matching= Matching(Onto, sentence);          list = element.parent;
    If(matching){                                list = element.child;
      url_queue(sentence.link);                  return list;
      check same link                        }
    }                                        Matching(Array a, string s){
}                                                if(s.word != 1){...}
                                                 while(a.size()){
                                                     .... }
                                             }
-
setIndex(element); 현재 Ontology의 연관된 단어
Matching(index, string); Ontology와 문장을 비교해서 매칭된 단어를
확인한다.
url_queue(element); 문장에 사용된 링크를 queue에 넣어 준다.
^
```

<그림 7. 문장 비교 알고리즘>

4.5 예외 경우
문장을 처리하는 과정에서 평가 대상에서 제외 되는 경우가 몇 가지 발생하게 된다.
첫 번째로 링크가 이미지로 표현 되어 있을 경우가 있는데, 이와 같은 경우에는 평가 대상에서 제외된다.
두 번째로 평가 할 수 있는 단어가 존재 하지 않는 경우로 웹 문서를 작성할 때, "여기", "here", "click"등과 같이 알 수 없는 의미의 숫자나 기호가 표시된 링크가 이것에 포함 된다.

5.결론
인터넷의 대중화로 인한 정보의 홍수 속에 웹 검색에 대한 중요도는 날이 갈수록 높아지고, 중요도가 높아짐에 따라 효율적인 crawler에 대한 필요성 또한 더욱 커지고 있다. 하지만 기존의 crawler로써는 다양화, 특성화 되어 가는 사용자의 요구에 대해 많은 한계가 있다. 따라서 이 논문에서는 이러한 한계를 해결하기 위해 하나의 문서에서 사용자에게 특성화된 주제를 찾아냄으로써, 사용자 취향에 맞는 문서를 제공할 수 있는 web crawler가 개발될 수 있을 것이다.
향후 사용자가 생성한 온톨로지를 이용할 수 있게 함으로써 좀 더 유연한 web crawler가 될 수 있을 것이다.

참고 문헌

[1]M. Ehrig, A. Maedche, "Ontology-Focused Crawling of Web Documents," In SAC2003, USA, 2003.

[2]S. Chakrabarti, M. van den Berg and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," In WWW-8 1999.

[3]F. Menczer, G. Pant and M. E. Ruiz, "Evaluating Topic-Driven Web Crawlers," In SIGIR'01, 2001.

[4]J. Cho, H. Garcai-Molina and L. Page, "Efficient Crawling Through URL Ordering," Computer Networks and ISDN Systems, 30(1-7):161-172,1998.

[5]M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori, "Focused Crawling Using Context Graphs," In VLDB-00, 2000.

[6]S. Sizov, J. Graupmann, M. Theobald, "From Focused Crawling to Expert Information: an Application Framework for Web Exploration and Portal Generation," In VLDB, 2003.