수식 관계를 이용한 키워드 추출을 통한 검색 과정의 효율성 향상

문욱성°, 이신목
대구고등학교°
한국과학기술원 전자전산학과 전산학전공
ukseong@gmail.com°, smlee@world.kaist.ac.kr

Keyword Extraction Using Modifying Relation to Improve Search Experience

Ukseong Moon°, Sheen-mok Lee
Daegu High School°
Division of Computer Science, KAIST

요 약

정보화 시대에 방대한 양의 정보에서 필요한 정보를 효율적으로 찾아내는 것은 그 무엇보다도 중요하다. 이를 위해 많은 검색 엔진이 효율적인 검색 결과 제공을 위해 노력하고 있지만 그 인터페이스의 문제로 인하여 사용자가 검색 결과를 효율적으로 받아들이기 어려우며, 또한 원하는 정보를 검색하기 위해서는 일정 수준 이상의 검색 능력을 필요로 한다. 이 논문에서는 기존의 검색 엔진의 인터페이스 변경을 통하여 시각적인 연관성 정보를 제공하며 이를 통해 사용자가 검색 능력에 구애 받지 않고 정확한 답을 얻을 수 있도록 유도한다. 또한 이 과정에서 기존의 키워드 추출 알고리즘의 문제점을 발견하여 이를 단어간의 수식 관계를 이용하여 해결하였다. 또한 단어간의 수식 관계를 이용하여 효율적으로 문서간의 연관성을 생성할 수 있는 알고리즘을 제시하였다.

1. 개요

정보를 효율적으로 받아들이고 정확한 검색 결과를 얻기 쉽도록 하는 것은 정보화 시대에서 검색엔진이 갖추어야 하는 중요한 요소이다.

현재의 검색엔진은 사용자의 검색 능력의 편차에 따라 원하는 결과를 얻을 가능성의 차가 큰 현상을 보인다. 또한 정보를 검색해줄 뿐 검색된 정보를 사용자가 받아들이는 과정에 대해서는 사용자와 정보 제공자가 전적으로 부담하여야 한다는 문제점이 있다.

하지만 검색 결과에 연관된 이미지를 표현하는 것[1]과 같은 부가적인 정보를 제공하는 방식으로 사용자의 검색 과정을 돕는 것에 대한 기존 연구들을 통해 사용자가 검색 능력에 구애 받지 않고 효율적으로 원하는 정보를 검색할 수 있도록 현재의 검색엔진을 발전시킬 수 있다고 생각하여 새로운 검색엔진 인터페이스를 제안한다.

검색에서의 최대 목표는 원하는 것을 정확하게 찾는 것이다. 이를 위해서는 각각의 정보에서 전달하고자 하는 내용이 무엇인지 분석하여 사용자가 관련된 내용을 검색할 해당 내용을 검색하여 제공하는 것이 필요하다.

이 논문에서는 문서의 내용을 표현하기 위한 방법으로 키워드를 이용하였다. 키워드는 문서의 내용을 효율적으로 표현할 수 있는 방법으로써 이와 관련해 이미 많은 연구에서 특정 정보의 내용을 효율적으로 표현하기 위한 수단으로 키워드를 이용하였다.[2,3] 하지만 기존 키워드 추출 알고리즘으로 추출된 키워드를 이용하여 문서간의 연관성을 생성하였을 때에 그 정확도가 현저히 떨어져 이 논문에서는 새로운 키워드 추출 알고리즘을 제안한다. 기존 알고리즘의 특징은 '출현 빈도수가 높은 것'을 키워드로 추출하는 것[3,4]이지만 이 논문에서 새롭게 제시하는 알고리즘은 출현의 빈도수가 아닌 문서 내 단어간의 수식 관계를 이용하여 각 문서에서 표현하고자 하는 핵심이 무엇인지 파악하였다.

또한 새로운 키워드 추출 알고리즘으로 만들어진 보다 정확한 문서간 연관성을 시각화하여 제공함으로써 사용자가 검색된 정보를 받아들이는 과정의 효율성을 향상하였다.

2. 현재의 키워드 추출 알고리즘

현재의 키워드 추출 알고리즘의 핵심은 단어 출현 빈도수이다. 특정 카테고리 내에서 출현하는 빈도수가 많은 단어, 특히 명사를 키워드로 추출하여 특정 내용을 표현하였다.[3,4] 하지만 빈도수를 이용하여 키워드를 추출하게 될 경우 문서의 내용을 분석하는 데에 한계가 있었고 이를 극복하기 위하여 사전 내에서의 상관 관계를 이용한 키워드 추출 알고리즘이나[5], 명사만을 사용하는 것이 아닌 명사와 동사간의 수식 관계를 이용하여 문서의 세부적 내용을 표현하려는 노력[6]이 있었다.

3. 새로운 키워드 추출 알고리즘

이 논문에서 키워드는 일반적인 내용을 담고 있는 문서 그룹에서 각 문서간의 연관성을 생성하기 위한 용도로 사용된다. 하지만 현재의 키워드 추출 알고리즘을 이용하여 문서간 연관성을 생성하였을 경

우 그 정확도가 낮아 새로운 키워드 추출 알고리즘을 제안하게 되었다.

새로운 키워드 추출 알고리즘의 핵심은 단어간의 수식 관계를 이용하는 것이다. 단어간의 수식 관계를 이용할 경우 각각의 문장에서 주요하게 다루는 단어가 무엇인지 파악할 수 있어 기존의 빈도수에 기반한 키워드 추출에 비해 문서의 세부적인 내용까지 고려할 수 있어 문서의 내용을 표현하기에 더 효율적이다.

3.1 키워드 추출의 목적

이 논문에서는 문서간의 연관성을 시각화하여 검색 시에 사용자에게 제공함으로써 검색 과정의 효율성을 높이는 것이 목적이다. 따라서 추출된 키워드는 문서간의 연관성을 효율적으로 비교할 수 있어야 한다.

3.2 단어간 수식 관계

문장 내에서의 단어간 수식 관계는 해당 문장에서 각 단어에 대한 어떠한 세부적인 정보를 주려고 하였는지 판단할 수 있게 한다. 출현 빈도가 높다는 것은 해당 단어의 쓰임이 많다는 것을 뜻하지만 수식을 많이 받고 있다는 것은 그만큼 문서에서 그 단어에 대한 세부적 정보를 많이 주려고 노력하고 있다는 것을 뜻하므로 핵심 내용을 표현하는데 더 적합하다. 따라서 새로운 키워드 추출 알고리즘에서는 문장의 주어를 중심으로 단어간의 수식 관계를 분석하여 수식을 많이 받는 단어에 대한 가중치를 부과해 키워드로 추출하였다.

3.3 추출 과정

문서의 단어간 수식 관계를 고려한 키워드 추출을 위해서 우선 문서의 모든 문장에 대하여 Connexor 파서(http://www.conexor.fi/)를 이용한 구문 분석을 실행한다. 구문 분석으로 얻어진 수식 관계를 토대로 주어를 중심으로 한 단어간 연관성을 구성한다. 단어간 연관성이 구성되면 각 주어들의 수식어를 탐색하여 각 주어가 수식 받은 횟수를 세고, 이 횟수가 평균을 넘어서면 키워드로 추출된다. 수식어를 탐색할 때에는 주어의 직접적인 수식어 외에도 간접적인 수식어까지 모두 탐색하게 되는데 이는 문서에서 간접적인 수식을 통하여 정보를 제공할 수도 있기 때문이다. 탐색을 할 때에는 주어 외에도 문장의 내용을 표현할 수 있는 명사와 기수 등도 포함하게 된다.

모든 단어간의 연관성을 분석하므로 기존의 키워드 추출 알고리즘들에서 명사만을 사용함으로써 겪게 되는 세부적 내용을 고려하지 못하는 문제를 해결하였다.

4. 문서간 연관성 생성

연관된 문서란 문서의 핵심적 내용과 개념이 같으나 그 개념에 대해 제공하는 정보의 범주나 종류가 다른 것을 뜻한다. 따라서 문서의 모든 내용을 분석하는 것보다 해당 문서의 핵심적 내용을 포함하고 있는 키워드를 이용하여 연관성을 생성하였을 때 더 효율적이다. 현재까지의 문서간 연관성들은 그 정확도가 낮아 문서 그 자체 외에도 웹 상의 링크[7]나 사람의 지성[8]을 필요로 하였다. 하지만 이렇게 부가적 정보를 이용하여 연관성을 생성할 경우 해당 정보를 제공할 수 있는 환경에서만 적용할 수 있다는 적용 범위의 한계가 발생하며, 그 연관성 또한 문서 그 자체에 기반한 것이 아니기 때문이 항상 적절하다고 할 수 없다.

이 논문에서는 3에서 제시한 새로운 키워드 추출 방식을 이용하여 문서간 연관성을 생성하였다. 그리고 이 연관성의 정확도를 검증하기 위하여 위키피디어의 See also로 연결된 내부 링크와의 일치도를 비교하였다. 그 결과 아래에서 제공하는 키워드 추출 API를 이용하여 추출된 키워드를 기반으로 만들어진 문서간 연관성보다 새로운 키워드 추출 방식을 이용한 방식이 약 85% 더 높은 정확도를 나타내었다.[그림 3]

4.1 연관성 추출 과정

연관성 생성 과정은 기준 문서의 키워드를 공유하는 비율을 통해서 생성하게 된다. 이 방식은 기존의 키워드 추출 방식과 새로운 키워드 추출 방식에 모두 똑같이 적용된다. 우선 추출된 키워드를 각 문서와 짝을 지어 데이터베이스에 저장한다. 그 뒤 기준 문서의 각각의 키워드를 포함하고 있는 다른 문서들을 찾은 뒤, 각 문서들이 중복된 개수가 기준점을 넘을 경우 기준문서와 연관된 문서로 추출된다.

Algorithm to extract keywords using modifying relation

```
EXTRACT_KEYWORDS( target_document )
01  words << all the words of a target_document
02  modify << 0
03  keywords << ∮
03  possibles << ∮
04  for each words
05    if word ∈ noun, abbreviation, or cardinal numbers then
06      add word to possibles
07      modify << modify + | modifiers of word |
08    end if
09  end for
10  for each possibles
11    if | modifiers of possible | > ( modify / | possibles | ) then
      //if number of modifiers are bigger than average number of modifiers
12      add possible to keywords
13  end for
14  return keywords
END EXTRACT_KEYWORS
```

[그림 1] 수식 관계를 이용한 키워드 추출 알고리즘 의사코드
수식 관계를 통해 문서에서 추가적인 정보를 많이 주려고 노력한 단어가 무엇인지 알 수 있으므로 핵심 단어를 효율적으로 판단할 수 있다.

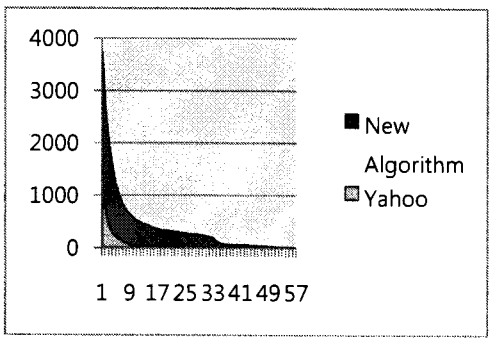
Algorithm to find related documents

```
GET_RELATED_DOCUMENTS( target_document )
01  keywords << EXTRACT_KEYWORDS( target_document )
02  word_contains << ∮
03  relations << ∮
04  for each keywords
05    word_contains <<
                word_contains + documents that contains keyword
06  end for
07  for each word_contains
08    if word_contains duplicated more than threshold in word_contains then
09      add word_contains to relations
10      end if
11  end for
12  return relations
END GET_RELATED_DOCUMENTS
```

[그림 2] 키워드를 이용한 연관 문서 추출 알고리즘
문서의 내용의 핵심을 나타내는 키워드를 통해 연관성을 생성하게 되므로 문서의 핵심적 내용이 비슷한 경우를 연관된 문서로 추출하게 된다.

4.2 검증

생성된 연관성의 정확도를 검증하기 위하여 위키피디어의 각 페이지간의 연관성을 기존의 키워드 추출 방식과 이 논문에서 제시하는 새로운 키워드 추출 방식에 기반하여 생성한 뒤 이 연관성을 위키피디어의 See also로 이루어진 링크와의 연관성을 비교하여 기존 연구와의 정확성을 비교하였다. 연관성을 비교할 때에는 각 방식에서의 모든 문서 중복 수치를 한번씩 기준점로 지정하여 전반적인 연관성의 정확도를 측정하였다.



[그림 3] 연관성 일치율 비교 그래프

X축은 기준점을, Y축은 일치하는 문서의 개수를 뜻한다. 야후에서 제공하는 키워드를 이용하여 생성된 연관성보다 이 논문에서 새롭게 제시하는 키워드 추출 방식으로 추출된 키워드를 이용하여 생성된 연관성이 약 85% 정도 더 높은 일치율을 보인다

5. 유사 문서

위키피디어와 같이 여러 도메인에 대한 방대한 양의 정보가 담겨있고, 각 페이지에서 중복된 내용을 가지지 않는 경우 위와 같이 핵심적인 내용을 기반으로 연관성을 생성하는 것이 효율적이다. 하지만 이를 뉴스와 같이 중복된 내용이 많고 특정 주제에 대한 다양한 글이 생성되는 환경에서는 핵심적 내용만을 이용하기 보다는 각 문서의 세부적 내용까지 고려한 유사한 문서를 찾아주는 것이 필요하다.

5.1 유사 문서 추출 과정

유사 문서를 추출하는 과정의 핵심은 각 문서 내에서의 단어간 연관성을 비교한다는 것이다. 단어간 연관성은 키워드 추출 과정에서 사용된 단어간 수식 관계를 이용하게 되는데, 이를 이용하여 두 문서간의 유사도를 비교하게 될 경우 문서에서 각 단어에 대해 제공하는 세부적 내용까지 비교하게 되어 유사한 문서를 찾기에 효율적이다.

유사 문서 추출을 위해서는 우선 비교 대상이 되는 문서를 추려낸다. 비교 대상 문서는 문서 내 단어 관계를 비교할 문서의 개수를 줄여주어 효율적인 유사 문서 추출을 가능하게 한다. 비교 대상 문서는 연관된 문서를 찾는 것과 똑같은 방식으로 찾게 된다. 비교 대상 문서가 추출이 되면 기준문서와 비교 대상 문서간의 단어간 수식 관계를 비교하게 되는데 이 비교의 시작점은 키워드이다. 따라서 유사 문서 추출 과정에서는 키워드를 '비교 시작점'이라 부른다. 비교 시작점을 시작으로 각 단어간의 수식 관계를 비교할 때에는 비교를 할 깊이를 설정할 수 있으며, 기본적으로는 각 수식어들이 일치하는 것을 찾게 되나 일치하지 않을 경우 WordNet에서의 유사도를 체크하게 된다. 일치하는 수식 관계의 비율이 기준점을 넘을 경우 유사 문서로 추출하게 된다.

Algorithm to compare modifying relations

```
CALC_MODIFY_SIMILARITY(compare_start_word, current_word, depth):
01  current_depth << 0
02  count << 0
03  entire_target_modifiers << 0
04  target_modifiers << modifiers of compare_start_word
05  target_num << | target_modifiers |
06  compare_modifiers << modifiers of current_word
07  while ( current_depth << depth ):
      //depth must bigger than 0
08    for each target_modifiers:
09      if target_modifier ∈ conjunctive, prepositional, or article then
10        add modifiers of target_modifier to target_modifier
11      else
12        entire_target_modifiers << entire_target_modifiers + 1
13        if target_modifier ∈ compare_modifiers then
            //find exactly same modifier
14          count << count + 1
15        else
16          for each compare_modifiers:
17            if compare_modifier has similarity with target_modifier in WordNet then
18              count << count + 1
19            end if
20          end for
21      end if
22    end if
23  end for
24  target_modifiers << modifiers of target_modifiers
25  current_depth << current_depth + 1
26  end while
27  return target_num ( count / entire_target_modifiers )
END MODIFYING_SIMILARITY
```

[그림 4] 문서간 수식 관계 비교 알고리즘

Algorithm to find similar documents

```
FIND_SIMILAR_DOCUMENTS( target_document, depth ):
01  similar_document_list << ∮
02  comparable_documents <<
          GET_RELATED_DOCUMENTS( target_document )
03  compare_start_words <<
          EXTRACT_KEYWORDS( target_document )
04  for each comparable_documents:
    //get each comparable document of a target document
05    probability << 0
06    current_words <<
            EXTRACT_KEYWORDS( comparable_document )
07    for each compare_start_words:
      //get each comparison start word of target document
08      if compare_start_word ∈ current_words then
        //find exactly same comparison start word
09        probability << probability +
            CALC_MODIFY_SIMILARITY( compare_start_word,
                                    current_word,
                                    depth )
10      else
11        for each current_words:
12          if current_word has similarity with compare_start_word in WordNet then
13            probability <<
                    probability +
                    CALC_MODIFY_SIMILARITY( compare_start_word,
                                            current_word,
                                            depth )
14          end if
15        end for
16      end if
17    end for
18    if probability > threshold then
19      add comparable_document to similar_document_list
20  end for
21  return similar_document_list
END FIND_SIMILAR_DOCUMENTS
```

[그림 5] 유사 문서 추출 알고리즘 의사코드

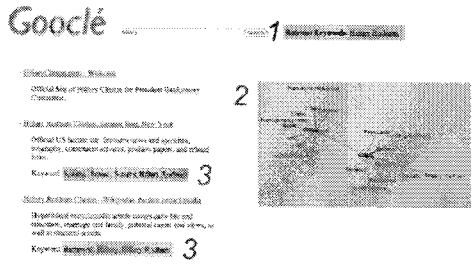
6. 검색 과정의 효율성 향상

현재의 검색 엔진은 결과를 랭킹하여 보여줄 뿐 검색 결과를 사용자가 효율적으로 받아들일 수 있도록 하고 있지는 않다. 하지만 기존의 몇몇 연구들에서 보여주듯이 사진과 같은 연관된 시각적 정보를 사용자에게 제공할 경우[1] 검색된 결과를 더 효율적으로 받아들일 수 있다.

이 논문에서는 새롭게 제시하는 키워드 추출 방식을 이용하여 검색 과정의 효율성을 향상시키려 하였다. 검색 엔진의 결과물에 키워드를 이용한 시각적 정보를 제공함으로써 사용자가 검색 결과를 일일이 분석하여 그 중 필요한 정보를 찾아내어 재구성할 필요 없이 각 검색 결과의 주요 내용과 연관된 정보를 한눈에 파악할 수 있도록 하여 정보를 쉽게 받아들이고, 그를 통해 정확한 답으로 사용자를 유도할 수 있도록 하였다.

6.1 새로운 인터페이스

이 논문에서 검색 과정의 효율성을 향상하는 주요 방식은 검색 엔진의 인터페이스를 사용자가 정보를 받아들이기 쉽도록 바꾸는 것이다. 이를 위해 Gooclé(Google과 불어에서 키워드를 뜻하는 moclé를 결합하였다)란 툴을 제작하여 검색 결과에서 페이지의 제목, 요약만을 제공하는 것이 아니라 각 검색 결과 페이지의 키워드 그리고 현재 검색어와 연관된 다른 키워드와 단어간 연관성을 시각화하여 보여줌으로써 사용자가 검색 결과를 읽고 중요한 내용을 찾아내어 재구성하는 과정을 없애 정보를 효율적으로 받아들일 수 있도록 하였으며, 시각화되어 제공되는 연관된 단어들을 통해 정확한 검색어를 모르거나, 잘못된 답으로 시작하더라도 정확한 검색어나 답으로 사용자를 유도해 나갈 수 있도록 하였다.

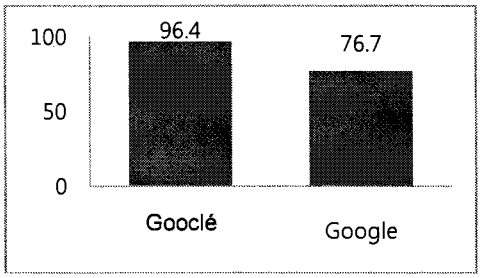


[그림 6] Gooclé에서의 새로운 인터페이스

1) 검색한 단어와 연관된 단어
2) 현재 페이지의 연관된 단어의 시각화된 연관성
3) 각 페이지 별 키워드

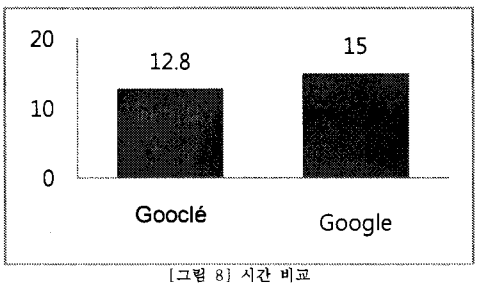
6.2 검증

연관성을 시각화하여 제공하는 새로운 인터페이스가 사용자에게 어떤 이익을 줄 수 있는지 검증하기 위하여 명확한 정답이 있는 문제 8개를 각 7명씩 두 그룹으로 나누어진 대학원생들에게 나누어주고 한 그룹은 구글을 통해서, 또 다른 한 그룹은 새로운 인터페이스가 적용된 Gooclé를 통해서 문제의 정답을 찾도록 하고, 각 실험 참가자들의 정답률, 모든 문제에 대답하는데 걸린 시간, 총 검색 횟수, 그리고 총 검색 횟수의 표준 편차를 비교하였다.



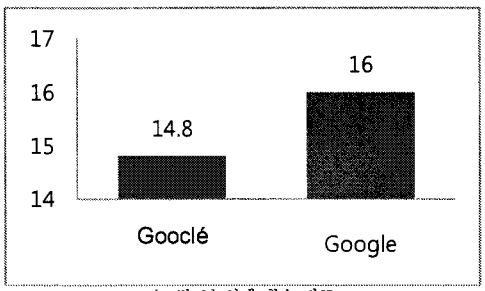
[그림 7] 정확도 비교

Google과 Gooclé을 사용하였을 때 Gooclé가 질문에 대한 정답률이 Google보다 약 20% 높다.



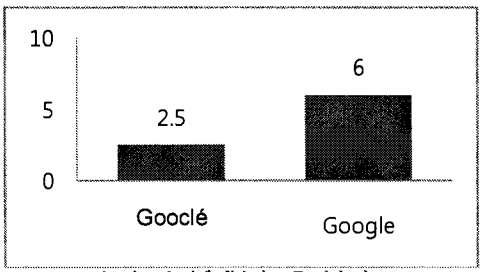
[그림 8] 시간 비교

Google과 Gooclé의 모든 질문에 답하는데 걸린 시간을 비교하면 Gooclé가 약 20% 빠르다.



[그림 9] 검색 횟수 비교

평균적 검색 횟수가 비슷하지만 시간은 오히려 더 짧아 시각화되어 제공되는 단어간의 연관성이 사용자를 정확한 답으로 빠르게 유도한다는 것을 알 수 있었다.



[그림 10] 검색 횟수의 표준 편차 비교

Google의 경우 검색 횟수의 표준 편차가 커 사용자의 검색 능력에 따라 올바른 검색 결과를 찾을 확률이 많은 차이가 나지만 Gooclé의 경우 그 질반에 해당하는 검색 회수의 표준편차가 생겨 사용자의 검색 능력에 덜 의존적이라는 것을 알 수 있었다.

6. 향후 연구

새롭게 제시한 유사 문서 추출 알고리즘의 검증과 연관 문서 알고리즘을 좀 더 검색에 알맞도록 변형하는 것이 필요하다.

7. 결론

본 논문에서는 문서간 연관성을 효율적으로 계산하기 위한 새로운 키워드 추출 알고리즘을 문서 내 단어간 수식 관계를 이용하여 추출하였다. 또한 이를 통해 키워드 추출에서 단순히 빈도수만을 이용함으로써 발생하는 세부적 내용을 고려하지 못하는 문제를 해결하였고, 문서 외적인 정보들을 이용하지 않고도 높은 정확도의 문서간 연관성을 생성하였다.

또, 단어의 수식 관계를 이용해 문서의 세부적 내용까지 분석할 수 있는 유사 문서 추출 알고리즘을 통해 단어 수식 관계의 또 다른 활용 방안을 제시하였다..

이러한 연관성들을 검색엔진에 시각화하여 적용하여 사용자의 검색 능력에 상관없이 원하는 내용을 시각화된 연관성을 이용해 쉽게 찾을 수 있도록 하였다. 이를 이용해 사용자의 검색 과정의 효율성을 전반적으로 향상시킬 수 있었다.

8. 참고문헌

[1]Zhao, R. & Grosky, W. "Narrowing the Semantic Gap - Improved Text-Based Web Document Retrieval Using Visual Features", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 4, NO. 2, 2002.

[2]Honda, T. et al. "Automatic Classification of Websites based on Keyword Extraction of Nouns", ENTER, 2006.

[3]Matsuo, Y. & Ishizuka, M. "Keyword Extraction from a Single Document using Word Co-occurance Statistical Information", International Journal on Artificial Intelligence Tools, 2003.

[4]Hulth, A. et al. "Automatic Keyword Extraction Using Domain Knowledge", 2001

[5]Woo, Y. et al. "Automated Keyword Extraction Using Category Correlation of Data", ICCSA 2006, LNCS 3981, pp. 224-230, 2006

[6]Lopez-Lopez, A. & Tapia-Melchor, M. "Automatic Information Extraction from Documents in WWW", CONILECOMP, 1998.

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

[8]Gyongyi, Z. et al. "Combating Web Spam with TrustRank", VLDB Conference, 2004.