모바일 환경에서 파일 검색 엔진을 위한 효과적인 방식

Effective Scheme for File Search Engine in Mobile Environments

조종근, 하상은
숭실대학교 컴퓨터학과, 인하대학교 컴퓨터공학과

Jong-Keun Cho(jkdang@empal.com), Sang-Eun Ha(1stpasa@gmail.com)

요약

본 논문에서는 파일 검색 엔진에 대해 모델링하고, 파일 검색의 정확도와 속도 향상을 위해 파일내의 내용들을 이용한 가중치 값 기반의 파일 검색 방식을 제안한다. 대부분의 파일 검색 엔진들은 빠른 검색 속도의 한계로 KMP와 같은 스트링 매칭 알고리즘을 사용해 왔다. 그러나, 이런 종류의 알고리즘들은 사용자가 원하는 파일들을 정확하게 찾아 주지는 못한다. 따라서, 모바일 환경에서 파일내의 내용들을 이용한 가중치 값 기반의 검색 엔진을 제안하고, 기존 방법들과 비교를 통해 제안한 방법의 우수한 성능을 증명한다.

■ 중심어 : | 문자 검색 | 검색 엔진 | 정보 검색 |

Abstract

This study focuses on the modeling file search engine and suggesting modified file search schema based on weight value using file contents in order to improve the performance in terms of search accuracy and matching time. Most of the file search engines have used string matching algorithms like KMP(Knuth - Morris - Pratt), which may limit portability and fast searching time. However, this kind of algorithms don't find exactly the files what you want. Hence, the file search engine based on weight value using file contents is proposed here in order to optimize the performance for mobile environments. The Comparison with previous research shows that the proposed schema provides better.

■ keyword : | Text Search | Search Engine | Information Retrieval |

1. 서 론

최근 구글과 네이버의 검색 엔진들은 네티즌 개개인이 인터넷 검색을 진행한 결과를 분석해 가장 많이 검색한 순으로 검색결과를 보여주는 '유저랭크 검색' 서비스를 시작하고 있다[1]. 하지만, 네티즌 개개인들이 원하는 문서를 찾기 위해서는 문서를 찾기 위해 입력한 키워드가 문서내에서 얼마나 빈번하게 나오느냐도 중요한 요인 가운데 하나다. 또한 키워드가 여러 차례 들어가는 문서는 키워드가 한번만 들어있는 문서에 비해 사용자가 찾기를 원하는 문서일 가능성이 높아질 수밖에 없다.

기본적으로 하나의 PPT 파일은 여러 개의 슬라이드로 이루어져 있고, 하나의 슬라이드는 다시 여러 개의

접수번호 : #081111-001
접수일자 : 2008년 11월 11일
심사완료일 : 2008년 11월 21일
교신저자 : 조종근, e-mail : jkdang@empal.com

텍스트로 이루어져 있다. PPT 파일 자료의 특성상 각각의 텍스트는 눈에 잘 띄고 간결하게 구성되어야 하기 때문에 텍스트가 포함하고 있는 내용에 따라 다양한 효과를 적용시키는 것이 일반적이다. 단순히 글자 크기에 변화를 주는 방법 이외에도 Bold, Underline, Color 등의 기본적인 강조효과들을 조합하여 하나의 텍스트를 너무 튀지 않게 표현하거나 강조할 수 있다. PPT 파일에 있어 주요한 의미를 가지는 키워드는 보통 제목에 상대적으로 커다란 폰트로 표시되거나 Bold 그리고 Underline 처리되어 나타나므로 이러한 특징들에 가중치를 주고 키워드와 일치한 패턴이 발견된 텍스트의 가중치를 검색결과에 사용하면 보다 지능적인 검색엔진을 구현할 수 있다.

본 논문에서는 PPT 파일내의 데이터 특성을 고려한 검색 알고리즘을 적용한 검색 엔진 및 구현하고, 기존 방법들과 비교를 통해 제안한 방법의 우수한 성능을 증명한다.

2. 시스템 설계

네티즌 개개인들이 인터넷 검색을 진행한 결과를 분석해 가장 많이 검색한 순으로 검색결과를 보여주는 '유저랭크 검색' 서비스가 등장했다. 검색포털 구글(Google)과 엠파스(Empas)는 사용자의 검색이용 행태를 분석한 후 이를 반영해 검색 정확도를 높인 '유저 랭크'(User Rank) 방식의 검색서비스를 시작했다[2]. 유저랭크 검색이란 사용자들이 많이 본 순서대로 검색 결과를 동적으로 재배치 해주는 신개념 검색 서비스이다. 이를 위해 엠파스는 1년 동안 사용자의 검색 데이터를 분석, 국내 최초로 자체 '동적 재배치 알고리즘'(DRA, Dynamic Relocation Algorithm)을 개발했다. 이를 이용해 엠파스는 지식, 블로그, 사이트, 뉴스, 이미지, 카테고리 등 포털 검색 서비스 이용시 가장 많이 검색한 정보 순으로 결과를 보여줌으로써 사용자들이 원하는 정보를 가장 빠르게 찾을 수 있도록 했다. 시기별, 이슈별로 변화하는 사용자의 관심도와 검색 이용률이 검색결과에 반영되기 때문에 같은 키워드라도 시기에 따라 검색 결과가 달라지는 방식이다.

그러나, 메타태그에 나타나는 키워드들은 여전히 고려해볼만한 대상이긴 하다[6]. 더블린 코어 메타태그에 포함된 일부 키워드들도 마찬가지다. 일부 검색엔진들은 키워드 메타태그를 완전히 간과하는 것은 아니지만 특히 그 키워드를 해당 페이지 다른 곳에서는 찾아볼 수 없을 때는 그 메타태그는 거의 무시한다. 이미지 보조 텍스트(ALT text, 그림이 보여지지 않을 경우 대치돼서 나타나는 텍스트)와 보조 텍스트의 관련성도 페이지 순위에 영향을 준다. 따라서 모든 이미지에 대해 보조 텍스트를 잘 만들어 달아 놓기 위한 시간 투자도 꼭 필요하다. 또한 페이지 순위에 있어서 키워드가 빈번하게 나오느냐도 중요한 요인 가운데 하나다. 키워드가 여러 차례 들어가는 페이지는 키워드가 한번만 들어있는 페이지에 비해 우선순위가 높아질 수밖에 없다. [그림 1]에서는 본 논문에서 제안한 PPT 파일의 검색엔진 시스템 구성도이다.

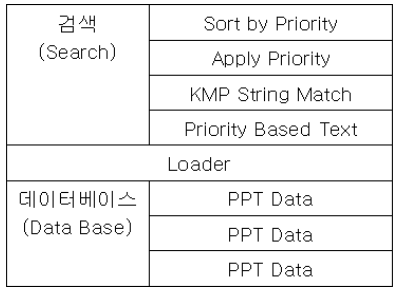


그림 1. 제안한 검색엔진 시스템

- PPT Data : 데이터베이스에 저장되는 데이터는 전처리를 수행한 데이터로 PPT 파일의 Property를 분석하여 미리 정의된 수식에 의해 계산된 가중치와 PPT 파일에서 뽑아낸 Plain Text 들을 저장한다.

- Priority Based Text : 데이터베이스에서 Load된 데이터는 Plain Text와 Text별 Priority 데이터를 포함한다.

- KMP String Match : KMP 알고리즘을 이용해 데이터베이스에서 읽어온 데이터 중 Plain Text 를

기준으로 사용자로부터 요청된 Key 를 검색한다.

- Apply Priority : KMP String Match를 통해 Key가 검출되면 Text별로 미리 계산되어진 Priority를 적용하여 해당 PPT 파일의 전체 가중치를 계산한다.

- Sort by Priority : PPT 파일별로 구해진 가중치를 리스트 하여 가장 높은 가중치를 갖는 PPT 파일의 데이터 별로 사용자에게 제공한다.

데이터베이스는 PPT 파일에서 추출한 Text 데이터와 각 Text의 글자 크기를 고려하여 계산된 가중치, 그리고 부가적인 PPT 파일에 대한 정보를 저장한다[4][5].

사용자로부터 검색 요청이 들어오면 시스템은 데이터베이스에 존재하는 모든 PPT Data를 검색하여 점수가 높은 PPT 파일순으로 정렬하여 응답하는데 작업의 순서는 [그림 1]의 아래쪽에 존재하는 레이어로 부터 상위 레이어로 올라가면서 진행된다.

데이터베이스에서 PPT 파일내 텍스트 데이터를 하나씩 읽어오는데 이때의 데이터는 우선순위가 포함된 데이터이다. 시스템은 UNICODE를 사용하며 KMP 패턴매칭 을 통해 일치하는 텍스트 패턴을 찾아낸다. 우선순위를 적용하여 검색결과를 도출하면 하나의 PPT 파일에 대한 검색결과가 된다. 검색작업은 데이터베이스에 존재하는 모든 PPT Data에 위와 같은 작업을 반복시키고 얻어진 결과를 취합해 가장 가중치가 높았던 PPT순으로 사용자에게 제공하게 되고. 이러한 방법을 통해 검색 결과에 대해 단순한 문자열 매칭보다 높은 만족도를 제공 할 수 있다.

- Query : 검색을 원하는 사용자가 입력한 Keyword는 별도의 처리없이 Search Engine으로 전송된다.

- PPT Property Analyze : PPT Data는 Plain Text 와 Multi Media Data로 이루어져 있다. 이 중 Text는 몇 가지 Property 를 가질 수 있는데 그 중에는 Size, Color, Bold, Italic 등이 있다. PPT 작성자가 특정 내용을 강조하기 위해 Text의 크기를 크게 한다거나 색을 조절하고 Bold 처리를 하는 등의 방법을 사용하는 점을 이용해 해당 효과가 적용된 Text의 경우 가중치를 상대적으로 높게 편성해 Plain Text와 가중치 정보를 분리하여 Data Base에 따로 제공한다.

- Data Base : DB는 Plain Text, Priority, Original PPT File 등의 정보를 포함하며 사용자가 검색을 원할 때 Key와 일치하는 문자열이 존재하는 Plain Text,와 해당 Text의 가중치를 같이 제공한다.

- KMP String Match : Data Base 에서 PPT Data를 읽어와 KMP 알고리즘을 통해 사용자가 입력한 Keyword 를 검색하여 검색 작업을 수행한다.

- Priority Calculation : Keyword가 검출된 Text는 DB에 함께 저장된 Priority 데이터를 토대로 가중치를 계산하게 된다.

- Sort Data : 다수의 PPT Data가 다양한 가중치를 가지고 요청된 Keyword를 포함하고 있을 경우 가중치가 높은 순으로 사용자에게 제공하기 위해 내림차순 정렬을 선행한다.

3. 스트링 검색 알고리즘의 개요

3.1 문자열 검색 알고리즘

1) KMP 알고리즘

KMP 알고리즘은 Kunth, Morris, Pratt라는 3인의 이름 머리글자를 따서 KMP이다. 문자열 패턴 매칭은 입력으로서 텍스트 T 와 패턴 P 가 주어진다. 둘은 문자열이고, 프로그램상에서는 문자의 배열로 표현될 수 있다. 즉, 텍스트에서 패턴이 나타나는 모든 위치를 찾아내는 것이다. 예를 들어 텍스트 abababc이고, 패턴이 aba라면 첫번째와 세번째 위치이다. 간단히 생각할 수 있는 KMP 알고리즘은 다음의 [표 1]과 같다.

표 1. KMP 알고리즘

```
텍스트 길이는 n으로 표시하고 패턴 길이는 m으로 표시

for( i = 0 ; i < 텍스트 상의 마지막 위치 ; i++)
{
    j=i,
    k=패턴의처음위치
   while
   (텍스트j번째 글자와 패턴의k번째 글자가 같은 동안)
   {
      j++,
      k++;
   }
    if(패턴의 끝을 지났으면 ) 발견
   else
       미발견  }
```

이 알고리즘의 시간 복잡도는 바깥쪽 루프는 최대 n번 반복되고, 안쪽 루프는 최대 m번 반복될 수 있다. 실제로 이렇게 걸리는 예는 aaaaaaaaaaaaaaa에서 aaab 찾기 물론 최악의 알고리즘 중에 하나이다. 그러나, 문자열 패턴매칭분야에선 이미 최적의 알고리즘이 나와 있고, 예는 [그림 2]와 같다[3].

- 입력 T : a b c a b c a b c d a b d a b b a
- 패턴 P : b c a b c d
- 답을 찾은 경우 a b c a b c d

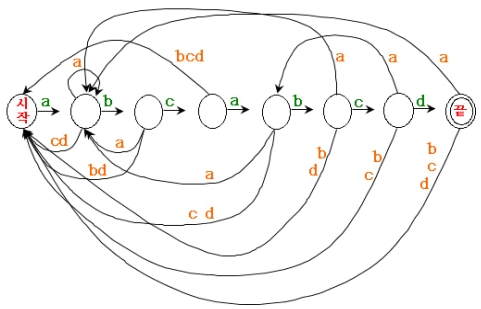


그림 2. KMP를 이용한 최적의 알고리즘 적용 예

알고리즘은 왼쪽의 시작 동그라미에서 시작한다. 알고리즘은 텍스트의 하나의 글자를 읽을 때마다 현재의 동그라미에서 그 글자가 붙어 있는 줄을 따라서 다음의 동그라미로 이동한다. 알고리즘을 시연해 보면 d를 읽은 직후 끝에 들어감을 알 수 있다. "끝" 동그라미에 들어가면 일치하는 패턴을 찾은 것이다. 이 알고리즘을 구현하기 위해서는 설명서를 만들고 표현하는 방법이 있어야 하지만 중요한 것은 이전 알고리즘과의 관계이다. 화살표를 따라 움직이는 과정은 이전 알고리즘과 대비할 수 있는데, 오른쪽 화살표를 따르는 과정은 이전 알고리즘의 안쪽 루프에서 한 단계 진행하는 것과 동일하다. 왼쪽 화살표를 따르는 것은 이전 알고리즘의 바깥쪽 루프 여러 개와 안쪽 루프 여러 개를 건너뛰는 것에 해당한다. 예를 들어 텍스트의 abcabca까지 읽은 상태를 보면, 이때 왼쪽 화살표를 따라서 이동하게 되는데 직후 비교되는 글자의 쌍은 텍스트의 8번째 글자인 b와 패턴의 5번째 글자인 b이다. 이것은 이전 알고리즘에서 2, 3번째 바깥쪽 루프와 안쪽 루프의 첫 4단계를 건너뛴 것이다. 이것이 가능한 이유는 패턴의 반복성 때문에 바깥쪽 2개의 루프에는 답이 있을 수 없고, 4번째 루프에서는 앞의 4글자가 같음을 알고 있기 때문이다. 이 정보는 패턴을 이용해서 얻을 수 있다.이 방법의 문제점은 설명서에 필요한 공간이 가능한 문자의 수 곱하기 패턴의 길이 정도로 길다는 점이다.

2) 보어-무어 기반의 검색 알고리즘

보이어-무어 알고리즘(Boyer-Moore string search algo- rithm)은 패턴이 텍스트에 정렬된 각 위치에서, 문자의 비교를 좌측에서 우측이 아니라 우측에서 좌측으로 수행해나가는 방법이다.

1. BM(Boyer-Moore) 알고리즘 패턴이 텍스트에 정렬된 각 위치에서 우측에서 좌측으로 문자를 비교함 불일치 문자방책, 일치접미부 방책을 사용하여 불일치 문자 방책과 일치 접미부 방책 중 패턴을 우측으로 이동하는 거리가 더 긴 것을 선택
2. 불일치 문자 방책(bad character heuristic) 텍스트에 있는 불일치가 발생한 문자가 패턴의 문자가 일치하도록 패턴을 오른쪽으로 이동
3. 일치 접미부 방책(good suffix heuristic) 패턴에서

불일치가 발생한 문자의 오른쪽에 있는 최대 접미부가 일치하도록 패턴을 오른쪽을 이동

4. 파일 검색엔진 알고리즘

기존 검색방법에 있어서 PPT 파일 검색을 위한 솔루션이 없었던 만큼 PPT Data의 특성만을 고려한 검색 알고리즘이 적용된 사례는 아직 공개된 바 없다. 일반적인 검색엔진에서 사용하는 방식은 단순히 문자열 데이터의 일치율을 바탕으로 이루어지는데, 이러한 방식의 경우 사용자가 원하는 문자열 자체의 검색에는 의미가 있으나 PPT Data를 검색하는 사용자의 의도에는 부합하기 힘든 경우가 많아지게 된다. PPT Data 검색을 원하는 사용자는 단순히 특정 키워드가 포함된 데이터가 아닌 키워드에 대한 내용을 원하는 경우가 많으며 이는 해당 키워드에 대한 상세한 설명이 담겨있는 발표자료, 강의교안 등의 자료에 부합한다. 좋은 발표자료, 혹은 강의교안에 있어서 반드시 특정 키워드가 많이 나오리란 법은 없으며 오히려 상세한 설명이나 프레젠테이션에는 해당 키워드가 별로 등장하지 않을지도 모르는 일이다. 기존 검색방법으로는 이러한 자료의 경우 상대적으로 낮은 검색율을 보일 수 밖에 없고, 이는 검색 결과의 품질 저하로 이어질 수 있다. 검색 키워드가 일치한 빈도수를 기준으로 일치율이 높은 Contents를 우선으로 검색할 때 특정 Contents의 우선순위 P를 수식 (1)처럼 표현할 수 있다.

$$P^{result} = N \qquad (1)$$

이와 같은 검색결과는 PPT 파일내에서 키워드가 여러 번 등장한 검색 결과를 우선시 하기 때문에 제목에 키워드가 존재하고 내용에는 키워드가 직접적으로 등장하지 않는 경우에 대해서는 검색결과를 보장할 수 없다. 이러한 부분에 대한 개선을 위해서는 PPT Data만의 특징을 이용해야만 한다. 하나의 PPT Data는 여러 개의 슬라이드로 이루어져 있으며, 하나의 슬라이드는 다시 여러 개의 텍스트로 이루어져 있다. PPT Data의 특성상 각각의 텍스트는 눈에 잘 띄고 간결하게 구성되어야 하기 때문에 텍스트가 포함하고 있는 내용에 따라 다양한 효과를 적용시키는 것이 일반적이다. 단순히 글자 크기에 변화를 주는 방법 이외에도 Bold, Italic, Underline 등의 기본적인 강조효과들을 조합하여 하나의 텍스트를 너무 튀지 않게 표현하거나 강조할 수 있다.

표 2. 글자크기, Bold, Italic 가중치 적용하는 알고리즘

```
 For( int i = 0 ; i < Number of Sizes ; i++ )
{
   Cnt = 0;
 For(  int j = 0 ; j < Number of Sizes ; j++ )
 {
   if( TextSizes[i] > TextSizes[j] )
   {
     Cnt++;     }
SizePriority[i] = 1+( 1.0 / Number Of Sizes )*Cnt;
}

BoldPriority_i = ( Text[i].Bold == true && # of Bold Texts
                   < # of Total Texts / 2 ) ? 1.414 : 1;
ItalicPriority_i = ( Text[i].Italic == true && # of Italic
    Texts < # of Total Texts / 2 ) ? 1.414 : 1;
```

PPT 파일에 있어서 주요한 의미를 가지는 키워드는 보통 제목에 상대적으로 커다란 폰트로 표시되거나 Bold로 표시함으로써 이러한 특징들에 가중치를 주고 키워드와 일치한 패턴이 발견된 텍스트의 가중치를 검색결과에 사용하면 보다 지능적인 검색을 구현할 수 있다. 본 논문에서는 PPT 파일에서 텍스트를 표현하는 여러 가지 방법 중 글자 크기와 Bold, Italic 에 초점을 맞추어 가중치를 계산하며, 알고리즘은 [표 2]와 같다.

위의 pseudo code는 텍스트의 폰트 크기, Bold, 그리고 Italic 별로 가중치를 구하는 알고리즘으로 이와 같은 방법을 통해 모든 텍스트 데이터에 각각의 가중치를 구해놓는다. 실제 검색 시에는 키워드가 일치한 텍스트의 가중치를 이용해 개별적인 가중치를 구하고 이를 누적시켜서 해당 PPT 자체의 우선순위를 구하는 방식을 취할 수 있다. 아래는 본 논문에서 제시하는 가중치 계산법의 수식 (2)는 다음과 표현된다.

$P^{result} = \sum(P_i^{text} * BoldPriority_i * ItalicPriority_i)$
(단, I=0이고, .Number of Texts 값이다) (2)

N이 i번째 텍스트에서 키워드가 발견된 빈도수를 나타내고 우변의 P가 폰트 크기로 구한 해당 텍스트의 가중치라고 할 때, 해당 PPT의 우선순위 P 는 가중치와 빈도수를 곱한 값에 Bold 와 Italic 의 적용 여부에 해당하는 가중치를 곱한 값의 누적된다. BoldPriority 와 Italic- Priority는 각각 Bold 와 Italic 처리된 텍스트의 가중치를 적용하기 위한 변수이며, Bold와 Italic은 각각 전체 텍스트 중에서 적용된 빈도수를 측정하여 전체 텍스트 대비 절반 이하의 텍스트만 적용되어 있을 때 해당 효과에 가중치를 적용하고, 그렇지 않을 경우에는 가중치를 부여하지 않는다. 각각의 텍스트는 슬라이드 내의 제목, 혹은 어떤 글 상자 안의 텍스트를 의미하며 이러한 텍스트들은 모두 고유한 가중치를 갖는다. 특정 텍스트 내에서 키워드가 발견되면 가중치 결과값에 해당 텍스트의 가중치를 더하고, 하나의 텍스트 안에서 키워드가 여러 번 발견 되더라도 해당 가중치가 여러 번 더해지는 방식이다. 위와 같은 수식에 의해 특정 경우엔 키워드가 발견된 빈도수가 상대적으로 적은 PPT Data가 보다 높은 가중치를 받고 검색결과의 상위에 존재할 수 있는데 이것은 앞서 설명한 PPT의 특성을 생각해 보면 문제가 되지 않는다. 단순히 빈도수가 많은 경우 해당 PPT 파일에서 사용자가 원하는 키워드에 해당하는 내용을 심도 있게 다루리라는 보장은 없으며 오히려 제목 등에 키워드가 존재 할 때 본문에는 키워드가 추가적으로 나타나지 않더라도 키워드에 대한 내용을 서술하고 있을 가능성이 크기 때문에 단순히 빈도수가 많은 경우보다 사용자가 원하는 PPT Data일 확률이 높다. 각 텍스트의 가중치는 여러 가지 방법을 통해 부여 할 수 있으며 이는 특정 주제, 또는 사용자에 따라 유동적일 수 있다. 예를 들면 특정 학회에서 주로 사용하는 발표자료의 양식에 따라 합리적인 가중치의 계산법이 달라질 수 있으며, 특정 사용자의 취향이나 PPT의 내용에 따라서도 역시 달라질 수 있다.

본 논문에서는 글자 크기와 Bold, Italic 에 의미 기반의 방법을 소개하였으며, 추후 보다 다양한 목적에 맞는 분석 방법을 도입한다면 더 높은 효과를 보일 수 있을 것이다.

특정 PPT Data에서 사용하고 있는 글자의 크기가 총 10가지라고 했을 때 상대적으로 큰 글자로 쓰여진 텍스트는 더 높은 가중치를 부여 받고 반대로 작은 글자로 쓰여진 텍스트는 낮은 가중치를 부여 받게 된다. 모든 PPT 파일에서 쓰이고 있는 단일 텍스트가 가질 수 있는 가장 높은 가중치를 1로 가정할 때, 1부터 10까지 10개의 글자크기를 사용하는 PPT 파일의 경우 글자 크기가 10인 텍스트의 가중치를 1로 선정하고 글자 크기가 1인 텍스트는 1/10 의 가중치를 부여 받게 된다. 마찬가지로 글자크기가2인 텍스트는 2/10, 5인 경우는 5/10과 같은 형태로 가중치를 부여 받으며 수식 (3)처럼 표현할 수 있다.

$$P^{text} = S^{text} / N \quad (3)$$

5. 실험 및 결과 분석

표. 3 검색 결과 비교표

	Hit Ratio (N=10)	Hit Ratio (N=100)	Hit Ratio (N=500)
Google	50%	25%	13%
Size Priority	100%	100%	72%
Bold Priority	100%	65%	57%
Italic Priority	50%	30%	28%
Integrated Priority	100%	100%	87%

[표 3]은 검색결과 중에 상위 20% 내에 Rank 되어 있는 자료들을 기준으로 일치율을 분석한 결과이며, Keyword 의 발견이 아닌, 상세한 정보제공 여부를 기준으로 일치 여부를 측정하였다. N개의 자료를 가진 데이터베이스를 기존의 알고리즘으로 검색하여 특정 Keyword를 포함한 PPT Data를 찾는다고 가정을 했을 때 10개의 자료 중 상위 20% 즉 순위로 2번째 내에 원하는 자료가 들어있는 확률은 50%로 측정이 되었다.

10개의 자료를 저장한 데이터베이스를 사용하는 일반 검색 사이트에서 상위에 리스트 되는 자료가 검색

요청자의 요구에 맞을 확률이 50%정도라고 했을 때. PPT 내부에서 글자 크기에 우선순위를 두어 그에 따라 검색결과를 조절할 경우 100%의 일치율 을 보였다. 기타 Bold와 Italic의 경우도 각각 100%와 50% 로 더 좋거나 같은 결과를 나타내었다. 실제 N개의 데이터 중 요청에 일치하는 자료는 2개에 불과했으며 해당하는 2개의 데이터가 각각 첫 번째와 두 번째의 일치율을 보이며 검색되는 것을 확인할 수 있었다. 100 개의 데이터를 가진 데이터베이스를 사용했을 때, 상위 20%에 리스트 되는 자료들 중의 일치율을 조사한 결과는 25%로 측정이 되었는데 10개일 때 에 비해 줄어든 수치는 자료가 많아질수록 검색 결과의 정확도를 더욱 보장하기 힘들어진다는 것을 의미한다. 반대로 글자 크기에 우선순위를 적용한 검색알고리즘은 여전히 100%의 일치율을 보였다.

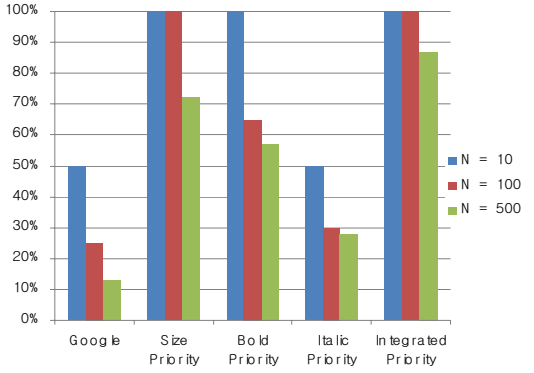


그림 3. 성능 측정 결과

Bold 와 Italic의 경우엔 자료의 개수가 많아짐에 따라 일치율이 조금 떨어지는 현상이 발생하였는데 이러한 점은 사용자 별로 PPT Data를 만드는 방식에 따른 차이라고 분석된다.

요컨대 500개의 자료를 가진 데이터베이스를 기준으로 Google의 검색 엔진과 비교한 결과 기존의 알고리즘이 13%, 글자크기 우선순위 방식이 72% 그리고 통합된 알고리즘을 사용하였을 때 87%로 PPT Data의 각 특성을 조합하여 사용하면 검색 품질을 더욱 높일 수 있음을 증명하는 결과를 [그림 3]에서 볼 수 있다.

6. 결론

본 논문에서 제안한 PPT 파일 검색엔진 방식은 기존 Google의 PPT 파일 검색 엔진보다 좋은 결과를 보였지만, PPT를 제작하는 사용자의 취향에 따라서 중요하거나 강조하고자 하는 문구에 Bold 속성을 추가 할 수도 있고 오히려 강조문구에는 Bold를 해제하고 글자 크기만을 크게 하거나 색을 변경하는 등의 방법을 이용하는 이용자도 있으므로 PPT의 제작 유형을 분석하지 않는다면 자료의 개수가 많아질수록 일치율이 떨어지는 현상이 발생함으로 향후 이를 개선하기 위한 연구가 필요하다.

참 고 문 헌

[1] http://www.seoworkshop.co.kr/seo-guide-book/SEO-ranking-guide.pdf

[2] N. L. Amy and D. M. Carl, "Google's PageRank: The Math Behind the Search Engine," Priceton university Press, Nol.03, pp.335-380, 2004.

[3] W. B. Michael and B. Murray, "Understanding Search Engines: Mathematical Modeling and Text Retrieval," SIAM Press, 2005.

[4] J. Zobel and A. Moffat, "Inverted files for text search engines," ACM, Vol.38, 2006.

[5] W, Mettrop and P, Nieuwenhuysen, "Internet search engines-fluctuations in document accessibility," Journal of Documentation, Vol.57, pp.623-651, 2001.

[6] S. Nobuyoshi, U. Minoru, S. Yoshifumi, and M. Hideki, "Fresh Information Retrieval Using Cooperative Meta Search Engines," information Networking, Vol.2344, pp.656-670, 2002.

[7] S. Narayanan and G. M. Hector, "Finding Near-Repicas of Documents on the Web," The World Wide and Database, Vol.590, pp.204-212, 1999.

[8] A. N. S. Craig and R. G. Gregory, "Connections: Using Context to Enhance File Search," CMU-PDL, Vol.05, pp.119-132, 2005.

[9] H. David, C. Nick, B. Perter, and G. Lathleen, "Measuring Search Engine Quality," Information Retrieval, Vol.04, pp.33-59, 2001.

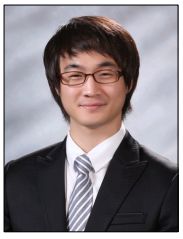
저 자 소 개

조 종 근(Jong-Keun Cho) 정회원

- 2001년 : 숭실대학교 컴퓨터학과(공학석사)
- 2004년 : 숭실대학교 컴퓨터학과(공학박사)
- 2004년 ~ 현재 : 모바일 3D 표준화포럼/전문위원
- 2004년 ~ 2006년 : (주)GOMID/수석연구원
- 2006년 ~ 2008년 : 삼성종합기술원 CTL
- 2008년 ~ 현재 : 삼성전자 TN총괄 통신연구소

<관심분야> : 모바일컴퓨팅, 멀티미디어, 컴퓨터그래픽스

하 상 은(Sang-Eun Ha) 정회원



- 2003년 ~ 현재 : 인하대학교 컴퓨터 공학과(공학사)
- 2005년 : 삼성종합기술원(계약직)
- 2007년 ~ 현재 : 삼성소프트웨어 멤버십

<관심분야> : 데이터베이스, 컴퓨터그래픽스