Journal of The Korea Society of Computer and Information
Vol. 20 No. 10, pp. 99-105, October 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.10.099

A Method to Measure the Self-Supplied News Volumes of Internet Newspaper Company

Dong-Joo Kim *, Won Joo Lee **

Abstract

The growth of internet infrastructure and a tremendous increment of internet users lead actively to found internet newspaper publishing companies, which are able to dig up and publish own news articles. In disregard of these quantitative growth of internet newspaper companies, the qualitative growth of them doesn't coincide with the quantitative growth. Therefore, to require social responsibility and to build healthy media environment, Korean government has put in force registration system of internet newspaper company. According to this system, internet newspaper companies have to produce at the inside over 30 percent of weekly publications, and this requisite increases the needs of its verification. This paper investigates technologies to measure the self-supplied news volumes of internet newspaper company, examines validity of them, and presents appropriate method to measure. To compare huge amount of news articles rapidly, the presented method is based on the modified edit-distance, which reflects human cognition of word and empirical information related with it. To prove correctness of our presented method, we show experimental results for some real internet news articles.

▸ Keyword : Internet Newspaper, Document Similarity, Edit-distance, Needleman-Wunsch Algorithm

I. Introduction

오늘날 인터넷 인프라의 놀랄만한 성장으로 인터넷 언론들의 수가 폭발적으로 증가하였으며, 웹2.0의 보편화와 모바일 플랫폼의 확대, 그리고 소셜네트워크서비스의 발달로 인한 인터넷 언론 환경은 크게 변화하고 있다. 특히 웹2.0의 보편화는 누구나 콘텐츠를 생산하여 자신만의 공간뿐만 아니라 대중을 상대로 공유할 수 있게 됨에 따라 블로그, 페이스북, 트위터 등의 소셜미디어를 바탕으로 하는 1인 미디어 시대가 열렸다.

90년대 초중반 시작된 인터넷 신문은 수익과 무관하게 오프라인 신문사의 인지도 강화와 부가서비스의 제공을 목적으로 초기에 신문사가 생산한 뉴스를 단순히 인터넷으로 게재하는 형태에 지나지 않았다. 그러나 2000년대 전후 인터넷 인프라의 성장과 이에 따른 사용자의 폭발적인 증가로 인하여 독자적인 기사를 발굴하여 게재할 수 있는 능력을 지닌 인터넷 신문 창간이 활발해지기 시작하였다. 즉, 단순히 종이 기사를 옮겨 인터넷으로 서비스 하는 종속형 인터넷 신문이 아닌 독립형 인터넷 신문의 급속히 증가하였다. 나아가 시공간 제약이 없고, 참여와 개방을 통한 높은 상호작용성을 강점을 바탕으로 하는 인터넷 미디어의 특성을 최대한 활용하여 개인, 시민 기자 등과 같이 사용자는 수동적인 수용자에만 머물러 있지 않고 자신의 의견을 능동적으로 표명하기 시작하였다.

이러한 인터넷 신문은 1995년 국내에서는 중앙일보가 처음으로 서비스를 시작하였고, 2014년 문화체육관광부의 '정기간행물 현황 등록 일람표'에 따르면 1997년에 18개에 불과하던 인터넷 신문은 성장을 거듭하여 오늘날 등록된 인터넷 신문사의 개수는 5,950개로 등록된 전체 정기간행물의 34%를 차지하기에 이르렀다. 인터넷 포털의 경우 국내에서는 1997년 야후에서 처음으로 뉴스서비스를 개시한 이후 뉴스서비스의 이용자가 급증하였고, 네이버, 다음, 네이트 등의 인터넷 포털은 사이트

• First Author: Dong-Joo Kim, Corresponding Author: Won Joo Lee
*Dong-Joo Kim (djkim@anyang.ac.kr), College of Liberal Arts, Anyang University
**Won Joo Lee (wonjoo2@inhatc.ac.kr), Dept. of Computer Science, Inha Technical College
• Received: 2015. 09. 08, Revised: 2015. 09. 24, Accepted: 2015. 09. 30.

당 약 40개가 넘는 언론사로부터 하루 약 5,000개 이상의 기사를 공급받아 500개 이상을 게시하고 있다.

인터넷 신문의 이러한 양적 성장에도 불구하고 질적 성장은 이에 부합하지 못하고 있는 실정이다. '신문 등의 진흥에 관한 법률'에 따른 여론집중도 조사위원회가 발표한 신문, 라디오, TV, 인터넷뉴스의 4대 매체에 대한 여론집중도 조사 결과 도달률(국내 인터넷 이용자 대비 해당 사이트 순방문자의 비중) 1% 이상인 뉴스 사이트는 24%에 불과하다. 그럼에도 불구하고 인터넷 뉴스는 즉시성과 전파의 확산 측면에서 상당한 영향력을 지니고 있으며, 신뢰성, 유해광고, 왜곡 보도, 과도한 경쟁에 따른 흥미위주의 자극적 기사 등의 문제에 시달리고 있다.

이에 따라 사회적 책임을 묻고 건전한 언론 환경을 구축하기 위해 정부는 2005년 7월 '신문등의자유와기능보장에관한법률'인 신문법을 개정하여 인터넷 신문의 등록제를 시행하기에 이르렀다. 시행령 3조에 따르면 인터넷 신문사는 독자적 취대 인력 2인 이상, 취재 및 편집 인력이 3인 이상이어야 하며 주간 게재 기사 건수의 30% 이상을 자체적으로 생산해야 한다. 물론 문화체육관광부는 현 시점에서 또다시 개정을 통하여 인터넷 신문 등록제는 강화하는 신문법 시행령 개정안을 입법예고하고 있다. 이 신문법 개정은 언론의 자율성을 위축시켜 건강한 언론 생태계를 해칠 우려가 있으며, 1인 미디어 시대에 역행하는 시대착오적 발상이라는 비판도 제기되고 있는 실정이다. 본 논문에서는 신문법 개정에 대한 정당성은 논외로 하고, 현 인터넷 신문사의 등록 요건 중의 기사의 자체 생산량을 측정하기 위한 기술을 다룬다.

본 논문은 인터넷신문 자체기사 생산량 측정을 위해 필요한 기술들을 조사하고 타당성을 검토하여 이에 적합한 기술을 제시한다. 이를 위해 2장에서 간단한 방법을 적용했을 때의 문제점을 제시하고, 문서 간 유사도를 측정하기 위한 기존 방법들을 나열하고 문제점을 분석한다. 문제점을 해결하기 위해 3장에서 제안하는 방법을 기술한다. 4장에서 제안하는 방법의 효율성을 검증한 뒤, 마지막 5장에서 결론을 맺는다.

II. Methods to Measure Similarity between Documents

1. Background

특정 인터넷 사이트에서 제공되는 인터넷신문기사는 자체 생산되는 경우, 다른 인터넷신문사로부터 제공받아 단순 유통하는 경우, 제공받은 신문기사를 통해 재생산하는 경우, 도용하거나 표절하는 경우로 나누어 볼 수 있다. 이들 중 법률에서 규정하고 있는 인터넷신문으로서 지위를 인정받기 위해서는 유통하는 경우와 재생산하는 경우, 도용하거나 표절하는 경우를 제외한 자체 생산되는 기사의 수를 측정할 수 있어야만 한다. 즉, 다른 어느 인터넷신문사에서도 생산되지 않은 고유의 자체 기사를 확인할 수 있어야만 한다. 본 논문은 이들을 정량화하여 기사들이 서로 얼마나 유사한지를 판단하는 것을 목표로 하고 있다.

신문기사에 대한 모방이나 도용, 표절이란 다른 사람의 기록물을 원저자의 동의 없이 전부, 혹은 일부를 도용하는 것인데, 이를 기계적인 방법으로 정확하게 정의하기는 어렵다. 개념적으로 쉽게 이해하기 위해 먼저 표절이나 도용의 판단을 위한 유사성 정도를 구분한다면 다음의 크게 다섯 가지 경우로 나누어 볼 수 있다.

① 완전히 동일한 경우
② 어느 한 기사의 내용이 다른 기사에 포함되어 있는 경우
③ 기사 내용의 일부가 동일한 경우
④ 주제는 동일하나 표면적 형태는 다른 경우
⑤ 완전히 다른 경우

일반적으로 ①, ②, ③의 경우를 표절로 간주되지만, ① 완전히 동일하거나 ⑤ 완전히 다른 경우를 제외한다면 나머지 ②, ③, ④의 세 가지 경우는 복합적으로 나타날 수 있기 때문에 그 구분이 매우 어렵다. 또한 ④의 경우에도 동일한 의미의 다른 단어로 대치하거나 조사와 어미의 활용 형태를 변경한다면 의도적 표절의 경우라도 판단하기가 어려워진다. 뿐만 아니라 ②나 ③의 경우에 문장이나 단락을 문맥의 흐름에 어긋나지 않게 교묘히 재배열하는 경우에도 표절의 탐지가 용이하지 않다. 또한 반대로 ②나 ③에서 제한적인 내용을 가져오되 출처를 명시한다면 표절에서 배제할 수 있어야만 한다.

일반적으로 신문 기사의 표절 검사를 위한 방법은 모든 신문기사 쌍을 서로 비교하되 비교 대상 단어의 표면적인 형태를 직접 비교하는 문자열 일치 방법(KMP, Boyer-Moore)을 사용하는 것이다. 이 방법을 사용할 경우 인터넷 신문사의 개수를 1,000개라고 가정을 하고, 각 신문사별 주당 게시 기사 수를 300개, 그리고 기사의 평균 길이를 1,024bytes라고 가정하면 비교 횟수는 식 (1)과 같이 $1.513\times10^{1,512,854}$회이다.

$$(1,000\times300)!\times1,024\approx 1.513\times10^{1,512,854} \quad (1)$$

따라서 1byte 문자 비교시간을 10^{-10}초라고 가정을 한다면 $4.77\times10^{1,512,851}$년이 걸린다. 이 시간은 허용할만한 시간이 아니지만, 이것이 전부가 아니다. 단순 문자열 일치 알고리즘은 완전 일치 문자열을 찾아내는 것을 목표로 하는 알고리즘이기 때문에 ② ~ ④와 같은 경우에는 적용할 수 없다. ② ~ ④와 같은 부분적으로 유사한 재생산되는 경우나 부분 표절 같은 경우에 대해 적용하려면 문자열 중간에 임의의 길이의 불일치, 혹은 대치, 삭제, 재배열 등의 문자열을 허용해야만 한다. 이럴 경우 비교 시간은 기하급수적으로 증가할 것이다.

따라서 이와 같이 다양한 경우에 대해 적응력 있게 빠른 시간 안에 모방이나 표절을 판단하기 위해서는 고전적인 단순한 문자열 일치의 방법을 넘어서는 기법들이 요구된다. 이에 적용 가능한 몇 가지 방법들을 살펴보고, 자체기사 생산량을 측정하기 위한 방법으로서의 타당성을 조사한다. 그런 다음 각 방법에

서의 몇 가지 장점을 취하여 자체기사 생산량 측정에 적합한 방법을 제시한다.

2. Information Retrieval Methods

기사의 독자 생산 여부를 판단하는 과정은 기준 문서를 다른 모든 문서와 일대일 비교를 수행한다는 점에서 정보검색과정과 동일하다. 정보검색은 용어, 혹은 색인어, 키워드, 등과 같은 정보 식별자(information identifier)로 각각의 문서를 특징 지워주고, 이 특징 정보의 비교나 특징 정보들 간의 거리를 측정하여 문서들 간의 유사성을 판별한다.

정보검색 방법에서 용어의 빈도에 기반한 벡터 공간 모형[1]의 첫 번째 문제점은 충분한 문서가 이미 수집되어 있어야만 한다는 것이다. 즉, 통계량으로서 가치 있는 량의 문서가 수집되지 않는다면 벡터 공간 내에서의 문서들 간의 일대일 유사도 거리는 다른 문서와의 상대적인 관계에서 무의미해진다. 충분한 문서가 수집되어야만 여러 문서들 간의 벡터 공간상의 거리가 가치 있어진다.

더욱 치명적이 문제점은 문서 내에서의 문맥을 반영하지 못한다는 것이다. 정보검색 기반 방법에서는 용어의 발생이 서로 독립적이라 가정하여 문서 내에 발생하는 용어 빈도만으로 문서를 벡터 공간 내에 사상하고 있다. 이러한 방법은 실제로 유사한 문서를 놓칠 가능성은 줄어들긴 하겠지만 본 논문에서 주요 쟁점이 유사하지 않은 문서를 구분하는 데는 매우 취약하다.

반면 정보검색 기반 방법의 장점은 표면정보에 민감하지 않다는 것이다. 이 장점은 앞서 기술한 독립 가정에 의한 빈도 기반 개념과 같은 것으로 장점이 되기도 하고 단점이 되기도 하는 문제이다. 즉, 정보검색 방법의 이러한 장점은 너무나 과도해 곧바로 단점이 되는 것이다. 따라서 정보검색 방법을 유사문서 판단에 활용하기 위해서는 문맥 정보나 표면 정보에 좀 더 민감하게 변경하는 것이 바람직하다.

또 다른 장점은 비교를 점진적으로 수행할 수가 있고, 매우 빠른 속도로 유사도를 계산할 수 있다는 것이다. 물론 유사도 계산을 위한 정보를 구축하는데, 즉, 색인어를 추출하여 역파일을 구성하는데 과도한 시간이 걸릴 수도 있기는 하지만 계산시 빠른 속도는 정보검색 기반 방법의 최대 장점이다.

3. Edit distance

신문 기사를 포함한 일반 전자 문서의 유사성 검사를 위한 보다 직접적인 편집거리(edit distance)라는 방법[2-6, 8]이 존재한다. 편집거리는 비교하는 두 문자열의 유사성 정도(엄밀히는 비유사성 정도)를 문자의 표면 정보를 직접 비교하여 측정하는 방법이고, 각 응용 분야에 따라 다양한 변형 알고리즘이 존재한다.

이 편집 거리 알고리즘의 근간을 이루고 있는 개념은 만약 일련의 행위의 집합이 항목열에 의해 기호적으로 주어질 수 있다고 했을 때, 두 기호열 표현을 직접 비교에 의한 두 행위열의 비교의 필요성은 행위들의 패턴을 규명하는데 매우 유용할 수 있다는데 있다.

편집 거리 알고리즘에는 몇 가지 다른 알고리즘이 있는데, 본 절에서는 편집거리 알고리즘을 살펴보고 편집거리 알고리즘의 유사도 평가 알고리즘을 사용하여 신문 기사 내용의 독자 기사 판단 방법을 소개한다. 본 절에서 소개하는 알고리즘은 기본적으로 비교 단위가 문자인데, 물론 문자를 비교 단위로 하더라도 문서 비교 문제에 적용하는데 어려움이 없지만 자연언어 문장에서 의미의 기본 단위인 단어나 형태소를 비교 단위로 삼는 것이 직관적으로 더 가치 있을 것이다. 따라서 이 장에서 특별한 언급이 없는 경우는 문자를 단어나 형태소로 간주하고, 문자열은 문장, 혹은 문서를 의미하게 된다.

3.1 Hamming Distance

Hamming 거리척도[2]는 통신 분야에서 데이터가 전송될 때 발생하는 오류를 발견하고 수정하기 위한 알고리즘이고, 오류를 평가하는 방법으로 고정길이의 이진 코드에서 비트의 뒤바꿈 수로 판단하였다. 즉, 동일한 길이의 두 문자열 사이의 Hamming 거리는 가장 단순한 방법으로 대응하는 서로 다른 기호들의 수이다. 즉, 비교 대상의 두 문자열에 대하여 한 문자열을 다른 문자열로 변경하기 위해 필요한 대치 연산의 수이다. 가능한 대치 방법이 여러 가지가 존재하겠지만 여기서는 최소 개수를 의미한다.

3.2 Levenshtein Distance

두 문자열의 유사도를 계산하기 위한 Levenshtein 알고리즘[3]은 패턴 인식, 생물정보학, 철자오류 교정, 음성인식 등 다양한 분야에서 사용된다. 이 알고리즘은 개념적으로 한 문자열을 비교 대상이 되는 다른 문자열로 변환하는데 필요한 삽입, 삭제, 교체 연산의 최소 횟수를 계산하는 알고리즘으로 값이 크면 클수록 비유사성 정도가 커지는 문자열 비교를 위한 비유사성 척도이다.

3.3 Damerau-Levenshtein Distance

두 문자열의 비교를 위한 Damerau-Levenshtein 거리 척도[4]는 워드프로세서에서의 철자 오류 교정을 위해 제안되었다. 이 척도는 Levenshtein 거리 척도와 동일하게 비교 대상이 되는 두 문자열에 대해 하나의 문자열을 다른 문자열로 변환하는데 필요한 삽입, 삭제, 교체 연산의 최소 개수를 의미한다. 그러나 Levenshtein 거리 척도와는 달리 한 문자열 내에서 이웃하는 두 문자의 교환 연산을 한 개의 연산으로 간주한다.

3.4 Sequence Alignment Algorithm

서열 정렬(sequence alignment) 알고리즘[5,6]은 생물정보학 분야에서 단백질이나 아미노 염기 서열에서 존재하는 부분 서열의 상관관계를 분석하는 방법이다. 서열 정렬의 목적은 관심 대상 서열과 상동성이 높은 서열들을 파악하여 서열의 기능을 추정하거나 관련 있는 서열들 간의 정략적 상관관계나 관련

기능 등을 예측하기 위한 것이다.

서열 정렬 알고리즘은 Levenshtein 거리 척도에서 삽입, 삭제 교체, 그리고 일치에 대한 가중치 집합을 일반화한 것이고, 엄밀한 의미에 비유사성(dissimilatiry) 척도인 Levenshtein 거리 척도와는 달리 서열 정렬 알고리즘에서 사용되는 척도는 유사성 척도이다. 따라서 Levenshtein 거리 척도에서의 각 연산에 대한 가중치는 해당 연산의 빈도수를 세는 역할만을 수행하지만 서열 정렬 알고리즘에서 삽입, 삭제, 교체 연산은 유사성 정도를 떨어뜨리는 역할을 하므로 벌점(penalty)이 주어지고, 일치에 대해서는 유사성 정도를 높이는 역할을 하므로 이점(advantage)을 주게 된다.

4. Dotplot

점도표(dotplot) 방법[7]은 원래 생물정보학학 분야에서의 연구가들이 DNA열의 자기유사성(self-similarity) 연구를 위해 시작한 방법으로 방대한 양의 디지털 정보에서 비교되는 두 문자열 매치의 패턴들을 시각화하기 위한 기술이다. 이 기술은 생물정보학 분야에서와 유사하게 일반 문서들 간에서도 유사한 서열(sequence)들을 찾아 전체적으로 유사한 부분을 쉽게 확인하기 위해 사용되고 있다. 문서들은 띄어쓰기 단위, 혹은 형태소 단위 등으로 분리하는 토큰화 과정을 수행한 뒤 각 토큰의 쌍들에 대해 쌍비교가 이루어진다. 토큰들이 일치하는 위치에 점을 찍거나, 혹은 유사성 정도를 나타내는 한가지의 색으로 점을 찍어 2차원의 점도표 공간을 구성하게 된다. 예를 들어, 만약 한 문서에서 세 번째 토큰이 또 다른 다섯 번째 토큰과 일치한다면 점도표 매트릭스의 (3, 5) 위치에 점 하나를 찍는 방식이다. 이렇게 완성된 점도표의 패턴은 사각형과 대각선의 시각적 이미지를 통해 해석된다.

III. The Proposed Method

지금까지 살펴본 여러 가지 알고리즘들과 기술들의 장단점을 토대로 구축할 수 있는 이상적인 시스템을 위해서는 탐색 범위 축소, 문맥의 반영, 표면 정보의 민감성 완화, 비교 횟수 축소 문제를 염두에 두어 두어야 한다.

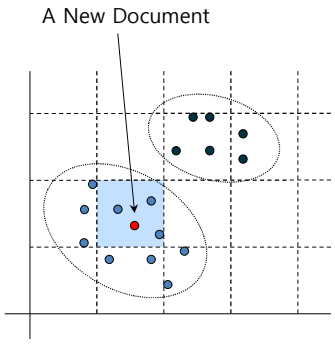

Fig. 1. Latticed Space

탐색 범위 축소를 위해 먼저 빈도기반의 정보검색 방법론을 적용하여 비교 대상이 되는 문서의 그룹화를 수행한다. 이때, 수행하는 그룹화는 그룹이 서로 최소한 30~50% 정도는 중첩되어야만 한다. 또한 각각의 그룹은 하나의 중심점을 갖는다. 비교하고자 하는 문서의 입력이 들어오면 Fig. 1과 같이 문서 벡터를 통하여 해당 문서와 가장 가까운 그룹의 중심점을 찾아낸다. 그 중심적이 소속되어 있는 그룹의 문서만을 대상으로 편집 거리 기반 방법을 적용하여 비교를 수행한다.

문맥정보를 반영하기 위해 연속 일치되는 부분은 길이에 비례하여 가중치를 높여주는 동적 가중치를 설정한다. 또한 표면 정보의 민감성을 완화하기 위해 형태소 분석된 어간/어근의 품사열의 유사도를 통합적으로 계산한다.

마지막으로 비교 횟수를 감소시키기 위해 대명사, 부사, 기호 등을 포함하고 문서의 의미에 큰 역할을 하지 않는 어절은 비교 대상에서 제외한다. 또한 어절 내에서도 형식 형태소를 제거하고 남은 어근/어간에 대해서 첫 문자와 마지막 문자만을 비교한다. 즉 '경찰은 쪽지가 발견된 뒤 수색작업을 벌이면서도'라는 내용에 대해 비교 대상 문자열은 '수색작업을'에서 형태소 분석된 '수색작업'과 '을'이라는 내용형태소와 형식형태소 중 형식형태소는 비교 대상에서 제외한다. 그리고 내용형태소 '수색작업'에서 첫 번째 문자 '수'와 마지막 문자 '업'이 비교 대상 문자열이 된다. 따라서 위 문장에서 비교 대상 문자열은 '경찰 쪽지 발되 뒤 수업 벌며'가 된다.

실험 평가를 위해 구현한 소규모 시스템에서 문자열 비교는 편집거리 방법들 중 Needleman-Wunsch 알고리즘[8]에 기반한 전역적 비교 방법을 사용하였다. 표면 정보에 대한 민감성을 완화하기 위한 표제어 정보나 품사 및 의미 정보와 같은 언어적 정보를 사용하지는 않았지만 단어(어절)의 경계 문자만을 비교하였기 때문에 민감성 완화뿐만 아니라 비교의 빈도수까지도 줄여 더 빠른 속도로 동작할 수 있었다.

사용한 Needleman-Wunsch 알고리즘은 식 (2)와 같으며 이 식에서의 가중치 값들은 $w_m = 1$, $w_s = 0$, $w_d = -1$로 설정하였다. 계산 결과 $D_{m,n}$에 대하여 순위화를 위한 일반화는 문서의 길이 의존성을 제거한 식 (3)에 의해 이루어졌다.

$$
\begin{aligned}
&D_{0,0} = 0 \\
&D_{i,0} = i \times w_d,\ D_{0,j} = w_d \times j \\
&D_{i,j} = \max \begin{cases} D_{i,j-1} - w_d & [\text{case1}] \\ D_{i-1,j} - w_d & [\text{case2}] \\ D_{i-1,j-1} - \alpha & [\text{case3}] \end{cases} \\
&P_{i,j} = \begin{cases} Diag & [\text{case1}] \\ Up & [\text{case2}] \\ Down & [\text{case3}] \end{cases} \\
&\text{where } \alpha = \begin{cases} w_m \text{ if } x_i = y_j \ (\text{match}) \\ w_s \text{ if } x_i \neq y_j \ (\text{mismatch}) \end{cases}
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
&SIM\ (X, Y) \\
&= \frac{D_{m,n} - w_d \times |m-n|}{w_m \times \min(m,n) - w_d \times |m-n|} \times 100(\%)
\end{aligned} \tag{3}
$$

따라서 완전히 동일한 두 신문 기사 X와 Y에 대한 유사도는 100이라는 결과가, 그리고 내용이 완전히 다른 두 신문기사, 즉, 공통적으로 같이 사용하고 있는 동일한 단어가 단 한 개도 없는 두 신문 기사의 경우 0이라는 값이 나올 것이다.

IV. Evaluations and Analysis

실험을 위한 데이터의 수집을 위해 규모(매출액)가 크고 기사 웹페이지에 포함되어 있는 부가 정보 데이터의 처리가 용이한 순으로 1개의 통신사를 포함하여 총 4개의 인터넷신문사를 선정하였다. 이 신문사로부터 2007년 12월 13일 11시 경에 속보로 올라온 기사를 무작위로 50개씩 총 200개의 기사를 수집하였다. Table 1은 4개의 인터넷 신문사로부터 추출한 기사의 수와 길이를 보여준다.

Table 1. the number and length of news collection

	Inc. Y	Inc. S	Inc. J	Inc. H	Total
# of news	50	50	50	50	200
avg. bytes	1,899	1,725	1,728	1,931	1,821
total bytes	94,950	86,250	86,400	96,550	364,150

Table 2는 수집된 신문기사 내용 중에 있는 인용정보를 토대로 수작업으로 자체 생산량을 조사한 것이다. Table 2는 임의로 수집한 기사에 대한 통계이므로 자체 생산 기사 수와 타사 인용 기사 수에서 타사 기사 전체를 인용한 기사가 수집되지 않아서 수집된 기사 집합 내에는 실제로 존재하지 않을 수도 있다. 즉 S사는 수집된 50개의 기사 중 28개의 기사를 Y사로부터 인용하였는데, 28개의 기사는 실제로 수집된 Y사의 기사 집합에 없을 수도 있다는 의미이다.

Table 2. Number and Length of Newspaper Collection

numbers within parenthesis are ratio (%)

		Inc.Y	Inc.S	Inc.J	Inc.H	total (except Inc. Y)	total (including Inc. Y)
# of self-producing news		50 (100)	20 (40)	12 (24)	24 (48)	56 (37)	106 (53)
citation of other Inc. news	Inc.Y	0	28 (56)	34 (68)	24 (48)	86 (58)	86 (43)
	Inc.E	0	1 (2)	0	1 (2)	2 (1)	2 (1)
	Inc.N	0	1 (2)	1 (2)	0	2 (1)	2 (1)
	Inc.P	0	0	3 (6)	0	3 (2)	3 (2)
	Inc.K	0	0	0	1 (2)	1 (1)	1 (0)

앞서 설명한 대로 Table 2의 기사 통계 중 타사인용 기사는 수집된 해당 타사 기사의 문서 집합에 포함되어 있지 않을 수도 있다. 시스템은 수집된 기사를 기준으로 동작하므로 타사인용 기사 중 수집된 해당 타사 기사의 문서 집합에 얼마나 포함되어 있는지를 알아야만 실험 후 시스템의 정확률을 파악할 수 있다.

Table 3. Number of news articles from other Incs

Inc. S	Inc. J	Inc. H	Total
8 (16%)	5 (10%)	9 (18%)	22 (15%)

수작업으로 조사한 결과 Table 3과 같이 수집된 신문사별 각각 50개의 기사 중 S사의 8개, J사의 5개, H사의 9개가 타사의 기사에 포함되어 있었으며, 모두 Y사의 기사였다. 수집된 기사 집합 내에서는 Y사를 제외한 모든 신문사의 기사가 Y사가 아닌 다른 두 신문사의 기사에는 포함되어 있지 않았다. 따라서 수집된 기사 집합을 기준으로 동작하는 시스템은 수집된 S사의 기사 50개 중 8개가 타사의 기사를 인용한 것으로 판정했을 때 S사의 기사에 대해 정확률 100%가 되는 것이다. 마찬가지로 수집 기사 총 150개 중 22개의 기사가 타사의 기사를 인용했다고 판단한다면 시스템의 정확률은 100%가 될 것이다.

수집된 인터넷 신문 기사 데이터를 제안하는 방법으로 얼마나 많은 기사가 자체 생산되지 않고 타사 기사를 인용한 것인지를 실험하였다. 먼저 Y사를 제외한 S사, J사, H사의 수집된 각각의 50개의 문서를 기준으로 다른 신문사 기사와의 유사도를 평가하였다. Table 4는 이에 대한 결과로 유사도 100%로 판정한 기사가 S사에 8개, J사에 5개, H사에 9개 있었고, 21~30%로 판정한 기사가 S사에 7개, J사에 8개, H사에 7개 있었다. 유사도 100%라고 판정한 각 신문사의 기사는 모두 예외 없이 수집된 Y사의 기사 집합 내에 존재하는 기사들이었다. 유사도 100%라고 판정한 기사는 Table 3에서 수작업으로 조사된 타사(Y사) 인용 기사수와 완전히 동일하므로 시스템의 정확률은 100%이다.

Table 4. The number of news by similarity

Newspaper Inc. / Similarity (%)	Inc. S	Inc. J	Inc. H	Total
100	8	5	9	22
51~99	0	0	0	0
41~50	1	0	0	1
31~40	0	1	0	1
21~30	7	8	7	22
11~20	9	11	8	28
1~10	7	9	9	25
0	18	16	17	51
Total	50	50	50	150

Table 4에서 주목할 만한 부분이 있는데, 타사 기사와의 유사도가 41~50% 범위 내의 있다고 판정한 S사의 기사 한 개와, 타사 기사와의 유사도가 31~40%의 범위 내에 있다고 판정한 J사의 기사 한 개이다. Fig. 2는 비교한 기사의 쌍인데, 유사도가 43%로 나왔다. 이 두 기사는 굵게 표시된 거의 절반에 가까운 내용이 완전히 동일한데도 불구하고 인용 표시가 없어 부분적이지만 표절이라고 볼 수 있다.

Inc. S	Inc. Y
제17대 대선이 일주일 앞으로 다가오면서 공식 선거운동이 종반전으로 접어든 가운데 주요 후보들은 각각 '대세 굳히기'와 '막판 대역전'을 목표로 마지막 승부수 준비에 들어갔다. 대통합민주신당 정동영 후보는 한나라당 이명박 후보의 BBK 관련 의혹을 끝까지 파고들고 범여권 단일화를 재시도해 막판 지각변동을 일으키겠다고 벼르고 있는 반면, 이명박 후보는 상대후보에 대한 네거티브 캠페인(폭로.비방전)은 지양하고 포지티브적 정책공약 캠페인에 주력, 남은 대선가도를 안전운행하겠다는 방침이다. 또 무소속 이회창 후보는 대북정책 등 안보 분야의 획기적 공약을 내세워 이명박 후보와 차별화함으로써 보수층을 총결집시키겠다는 구상이다. 신당 정동영 후보는 12일 민주화운동가인 고 지학순 주교가 봉직했던 강원도 원주 원동성당에서 기자회견을 갖고 "지금 전국적으로 이명박 후보에게 면죄부를 :	**제17대 대선이 일주일 앞으로 다가오면서 공식 선거운동이 종반전으로 접어든 가운데 주요 후보들은 각각 '대세 굳히기'와 '막판 대역전'을 목표로 마지막 승부수 준비에 들어갔다. 대통합민주신당 정동영 후보는 한나라당 이명박 후보의 BBK 관련 의혹을 끝까지 파고들고 범여권 단일화를 재시도해 막판 지각변동을 일으키겠다고 벼르고 있는 반면, 이명박 후보는 상대후보에 대한 네거티브 캠페인(폭로.비방전)은 지양하고 포지티브적 정책공약 캠페인에 주력, 남은 대선가도를 안전운행하겠다는 방침이다. 또 무소속 이회창 후보는 대북정책 등 안보 분야의 획기적 공약을 내세워 이명박 후보와 차별화함으로써 보수층을 총결집시키겠다는 구상이다.** 이런 가운데 **신당 정동영 후보**와 한나라당 이명박 후보는 이날 공교롭게도 강원도와 충북지역에서 일정한 시차를 두고 거리유세를 갖고 치열한 표심잡기 :

Fig. 2. Two news of similarity 43%

Inc. J	Inc. Y
중국 경제는 내년에 고도성장을 지속할 가능성이 크지만, 중국 진출 해외 기업들의 경영환경은 나빠질 것이란 전망이 **나왔다. 삼성경제연구소는 12일 '2008년 중국 경제에 대한 8가지 질문' 보고서에서** 내년은 베이징 올림픽이 열리는 해로, 중국이 글로벌 경제강국으로 진입하는 분기점이 될 것이라고 전망했다. 정상은 수석연구원은 "중국은 올림픽을 계기로 경제의 대외 영향력 확대를 노린다"며 "한국이 중국의 산업 고도화를 기회로 활용해야 한다"고 말했다. 다음은 보고서 개요. 중국이 2001년 세계무역기구(WTO)에 가입한 뒤 줄곧 연 **10%가 넘는 고도성장**을 해왔다. 하지만 **소득격차 확대, 환경 파괴, 물가 상승 등의 부작용**도 만만치 않았다. **중국의 경제성장률은 올해 11.4%에 이어 내년에도 10.7%를 이어갈 전망이다. 고용창출, 사회 안정, 낙후 지역 개발 등을** 추진하려면 여전히 고성장 동력이 필요하기 때문이다. **중국발** 글로벌 **인플레이션의 발생 가능성도 희박한 것으로** 보인다. **최근 중국의 물가급등은 식료품 가격 상승에 따른 것**이지, **서비스나 공산품 가격이 전반적으로 오른 것은 아니**다. **내년도 소비자물가 상승률 또한 4.2%로 전망돼 글로벌 인플레이션을 유발할 정도는 아니**다. 주식과 부동산 등 자산가격의 붕괴 가능성도 크지 않다. 버블이 낀 건 사실이지만, 자산가격이 내년에 조정기에 접어들면서 상승세가 둔화하는 수준에서 그칠 전망이다. 위안화 절상도 급격하게 이뤄지지 않을 것 같다.선진 7개국(G7)의 위안화 절상 압력은 거세지만 경제 안정을 중시하는 중국 정부의 의지가 확고하기 때문이다. ⋮	**중국은 내년에 고도성장을 지속할 가능성이** 높으며 인플레이션이나 자산가격 버블 붕괴의 위험도 크지 않다는 의견이 **나왔다. 삼성경제연구소는 12일 '2008년 중국경제에 대한 8가지 질문' 이라는 보고서에서** 중국경제가 글로벌 경제강국이 될 지 여부를 판단하려면 내년에 고성장 정책을 지속할 것인 지를 따져봐야 한다고 밝혔다. 그동안 중국은 **10%가 넘는 고도성장을** 하면서 **소득격차 확대, 환경파괴, 물가상승 등의 부작용**이 생겼다는 것이다. 특히 작년도 지니계수는 0.47로 위험수준인 0.4를 넘었고 소비자 물가 급등으로 공산당의 핵심기반인 농민과 근로자들 사이에 불만이 팽배해 있다고 연구소는 전했다. 연구소는 그러나 **중국의 경제성장률은 올해 11.4%에 이어 내년에도 10.7%의 높은 수준이 예상된다고 밝혔다. 고용창출, 사회안정, 낙후지역 개발 등을** 위해서는 아직도 고성장이 절실하기 때문이라고 설명했다. 성장률을 낮추기 위해 강력한 긴축정책을 펼 경우에는 고용시장이 붕괴되고 사회가 혼란스러워지는 문제가 발생할 것이라고 연구원은 진단했다. 연구원은 **중국발 인플레이션 발생 가능성도 희박한 것으로** 판단했다. **최근 중국의 물가급등**은 돼지고기 등 **식료품 가격 상승에 따른 것**으로 **서비스나 공산품 전반의 가격이 오르는 것은 아니**라는 설명이다. 또 임금상승에도 불구하고 생산성 향상을 감안한 단위노동 비용은 오히려 하락하고 있으며 에너지와 생산요소 가격에 대한 정부 통제도 지속되고 있다고 전했다. 중국의 **내년도 소비자물가 상승률은 4.2%로 예상되며 이는 글로벌 인플레이션을 유발할 정도는 아니**라는 것이 연구소의 분석이다. ⋮

Fig. 3. Two news of similarity 31%

또한 Fig. 3과 같은 기사 쌍은 유사도 31%로 판정한 것인데 기사 내용 중간에 부분적으로 완전히 동일한 내용이 존재하여 표절로 의심 받을 만한 쌍이지만, 두 기사 모두 외부의 특정 기관의 연구 발표 내용을 인용하는 기사이기 때문에 동일한 부분이 많을 수밖에 없는 경우이다.

V. Conclusions

본 논문에서는 인터넷기사의 자체 생산량을 측정하기 위한 방법들을 조사하고 타당성을 검토하였다. 이들 방법들 중 본 논문에서는 탐색범위를 축소하기 위해 빈도를 이용하여 전혀 유사하지 않은 문서를 걸러내었다. 또한 표면 정보의 민감성을 완화하기 위해 어근/어간의 품사열 비교를 수행하였으며, 비교 횟수의 감소를 위해 문서에 큰 영향을 주지 않는 대명사, 부사, 기호의 비교를 제외하였으며, 단어의 경계문자만을 비교하였다. 이들 정보와 비교 기준을 통하여 최종적으로 비교를 수행하기 위한 알고리즘으로는 Needleman-Wunsch 알고리즘에 기반한 전역적 비교 방법을 사용하였다. 실험결과 자체 생산된 기사를 엄격히 분간할 수 있었을 뿐만 아니라 부분적으로 유사한 내용의 기사도 알아낼 수 있었다.

그러나 본 논문에서 실험을 위해 사용된 기사는 주요 몇몇 신문사였으며, 특정 시간대에 게시된 매우 작은 량을 기사였을 뿐이다. 또한 추출한 기사는 무작위로 추출하였기 때문에 논문에 언급된 인터넷신문사들의 자체 생산량을 반영한다고 볼 수 없다. 따라서 향후 좀 더 광범위한 데이터 수집을 통한 장기적인 평가가 필요할 것이다.

REFERENCE

[1] G. Salton, et al., "A Vector Space Model for Automatic Indexing," Communication of ACM, Vol. 18, No. 11, pp. 613-620, Nov. 1975.

[2] Richard W. Hamming, "Error Detecting and Error Correcting Codes," Bell System Technical Journal, Vol. 29, No. 2, pp. 147-160, April 1950.

[3] V. L. Levenstein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reverables," Soviet Physics-Coklady Akademii Nauk SSSR, Vol. 163, No. 4, pp. 845-848, August 1965.

[4] F. Damerau, "A Technique for Computer Detection and Correction of Spelling Errors," CACM, Vol. 7, No. 3, pp. 659-664, March 1964.

[5] S. Altschul, "A Protein Alignment Scoring System Sensitive at all Evolutionary Distances," Journal of Molecular Evolution, Vol. 36, No. 3, pp. 290-300, March 1993.

[6] T. smith and M. Waterman, "Identification of Common Molecular Subsequences," Journal of Molecular Biology, Vol. 147, pp. 195-197, March 1981.

[7] K. Church and J. Helfman, "Dotplot: a Program for Exploring Self-Similarity in Millions of Lines of

Text and Code," Journal of Computational and Graphical Statistics, Vol. 2, No. 2, pp. 153-174, June 1993.

[8] S. Needleman and C. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins," Journal of Molecular Biology, Vol. 48, pp. 443-453, May 1970.

Authors

Dong-Joo Kim received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Hanyang University, Korea, in 1996, 1998 and 2007, respectively.

Dr. Kim joined the faculty of the College of Liberal Arts at Anyang University, Gyeonggi-do, Korea, in 2008. He is currently a Professor in the College of Liberal Arts, Anyang University. He is interested in natural language processing, opinion mining, critiquing system, machine learning, information retrieval.

Won Joo Lee received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Hanyang University, Korea, in 1989, 1991 and 2004, respectively.

Dr. Lee joined the faculty of the Department of Computer Science at Inha Technical College, Incheon, Korea, in 2008, where he has served as the Director of the Department of Computer Science. He is currently a Professor in the Department of Computer Science, Inha Technical College. He has also served as the Vice-president of The Korean Society of Computer Information and the Editor-in-Chief for the Journal of The Korean Society of Computer Information. He is interested in parallel computing, internet and mobile computing, and cloud computing.