Proceedings of KFIS 2001 Fall Conference, 2001. 12. 1

문서 길이 정규화를 이용한 문서 요약 자동화 시스템 구현

Implementation of Text Summarize Automation Using Document Length Normalization

이재훈 · 김영천 · 이성주
조선대학교 전자계산학과
Jea-Hoon Lee and Young-Cheon Kim and Sung-Joo Lee
Dept. Computer Science. Chosun University
E-mail : nuridepo@cafe.chosun.ac.kr, yckim@stmail.chosun.ac.kr,
sjlee@mail.chosun.ac.kr

ABSTRACT

With the rapid growth of the World Wide Web and electronic information services, information is becoming available on-line at an incredible rate. One result is the oft-decried information overload. No one has time to read everything, yet we often have to make critical decisions based on what we are able to assimilate. The technology of automatic text summarization is becoming indispensable for dealing with this problem. Text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user or task.

Information retrieval(IR) is the task of searching a set of documents for some query-relevant documents. On the other hand, text summarization is considered to be the task of searching a document, a set of sentences, for some topic-relevant sentences.

In this paper, we show that document information, that is more reliable and suitable for query, using document length normalization of which is gained through information retrieval. Experimental results of this system in newspaper articles show that document length normalization method superior to other methods use query itself.

Key Words : 정규화, 문서 요약 자동화

I 서론

인터넷과 같은 정보유통시스템의 발달로 인해 정보의 양은 하루가 다르게 지속적으로 증가하고 있다. 이러한 대량의 정보들 중에서 사용자가 원하는 정보를 찾아내기란 쉽지 않은 일이다. 누군가 각 정보의 내용을 요약해서 제공해 준다면 원하는 정보를 신속하게 찾아낼 수 있을 것이다. 이처럼 정보를 습득하는 과정을 원활히 하기 위해서는 정보에 대한 요약 작업이 필요하며, 방대한 정보의 요약 작업을 수동으로 하는 것은 비효율적인 일이 되어 자동 요약 시스템의 필요성이 대두되고 있다.

방대한 정보들 중에서 원하는 정보를 찾기 위해서 정보 검색 시스템을 사용하지만 정보 검색 시스템이 제시하는 검색 결과는 사용자가 하나씩 읽어보면서 확인하기에는 너무 많은 양이다. 이러한 정보 과적재(information overload) 문제는 정보 검색 시스템에서 해결 해야할 과제로 남아 있다.

일반적인 정보 검색 시스템들은 문서의 제목과 앞부분을 약간 보여주어 이 문제를 해결하려 하지만, 이 정도의 정보는 사용자가 검색 결과 문서의 적합성을 판단하기에 부족하다. 반면, 문서 요약 자동화 시스템에 의해 생성된 요약을 추가한 정보 검색 시스템은 사용자가 원하는 정보를 찾아내는데 시간을 단축시킴으로써 정보 과적재 문제에 대한 효과적인 해결책을 제시해줄 수 있다[3].

본 논문은 질의를 이용하여 정보 검색 시스템을 통해 검색된 문서를 정규화하여 질의에 대한 가중치를 산출하고 또한 문서 자체 신뢰도와 적합도가 높

은 문장을 선택하여 그 문서의 요약을 추출하는 시스템을 제안하였다.

본 논문은 다음과 같이 구성되었다. 2장에서는 문서 요약 자동화와 문서 길이 정규화 분야에 수행되었던 관련 연구들을 고찰하여 간략히 설명하였다. 3장에서는 본 연구에서 제안한 문서 요약 자동화 시스템에서 사용되는 요소들을 제안하고 4장에서 시스템을 실험하고 분석하였고 5장에서 결론에 대해 언급하였다.

II 관련연구

문서 요약이란 문서의 기본적인 내용을 유지하면서 문서의 복잡도, 즉 문서의 길이를 줄이는 작업이다[2]. 요약(Summarization)이란 원 문서에서 중요하다고 여겨지는 것을 선택하고, 혹은 일반화하는 방식으로 내용을 축소, 변형하는 작업이다. 이러한 요약에서 질의를 이용하는 것은 중요하다.

길이가 긴 문서는 일반적으로 동일한 단어(term)들의 반복적인 출현과 또한 여러 개의 서로 다른 단어들이 나타나기 때문에 길이가 짧은 문서에 비하여 질의와 높은 유사도를 나타낼 뿐만 아니라 검색될 가능성도 그만큼 높아지게 되므로 문서 길이 정규화는 대단히 중요하다[4].

2.1 문서 길이 정규화

문서의 길이 정규화(Document Length Normalization)는 다음 두 가지 이유로 인하여 대단히 중요하다. 첫째, 길이가 긴 문서는 일반적으로 단어(term)들이 반복적으로 나타나기 때문에 길이가 짧은 문서에 비하여 비교적 질의와 높은 유사도를 나타낸다. 둘째, 길이가 긴 문서는 일반적으로 여러 개의 서로 다른 단어들이 나타나기 때문에 길이가 짧은 문서에 비하여 질의와 높은 유사도를 나타낼 뿐만 아니라 검색될 가능성도 그만큼 높아지게 된다.

결국, 문서의 길이를 정규화를 하는 작업은 문서의 길이에 따른 유사도 계산 및 검색 가능성의 차이를 최대한 제거하여 검색의 형평성을 보장하기 위한 것이라 할 수 있으며, 대표적인 정규화 방법으로는 코사인 정규화(Cosine Normalization), 최대 단어빈도 정규화(Maximum tf Normalization), 바이트 길이 정규화(Byte Length Normalization)가 있다[4].

1) 코사인 정규화

Cosine 정규화는 벡터 공간 모델에서 가장 일반적으로 사용되는 정규화 방법이다. Cosine 정규화 공식은 다음과 같은 식(1)으로 산출할 수 있다.

$$C(S)=\sqrt{\omega_1^2+\omega_2^2+\cdots+\omega_n^2} \qquad (1)$$

여기서 ω_i는 단어의 문서에서 단어 출현 빈도(TF : Term Frequency)와 역문서 출현 빈도(IDF : Inverse Document Frequency)의 곱으로 문장에서 단어의 가중치를 나타낸다[5, 6].

2) 최대 단어 빈도 정규화

또 다른 일반적인 정규화 방법으로 이 방법은 단어 빈도(tf)를 그 문서에서 가장 많이 출현한 단어의 단어 빈도로 정규화시키는 방법이다.

SMART 시스템에서 tf 가중치 산출식은 아래의 식(2)이고, INQUERY 시스템의 tf 가중치 산출식은 식(3)이다.

$$0.5+0.5\times\frac{tf}{\max tf} \qquad (2)$$

$$0.4+0.6\times\frac{tf}{\max tf} \qquad (3)$$

3) 바이트 길이 정규화

최근 Okapi 시스템이 TREC에 참가하면서 제안한 정규화 방식으로, 문서의 바이트의 크기에 따른 정규화이다[4].

2.2 문서 요약

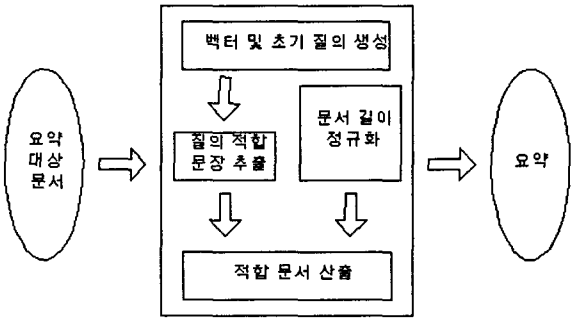
문서 요약이란 문서의 기본적인 작업을 유지하면서 문서의 복잡도, 즉 문서의 길이를 줄이는 작업이다[2]. 요약(summarization)이란 원 문서에서 중요하다고 여겨지는 것을 선택하고, 혹은 일반화하는 방식으로 내용을 축소, 변형하는 작업이다. 이러한 요약에서 질의를 이용하는 것은 중요하다.

문서 요약 자동화에 관한 연구들은 방법론에 따라 언어학적 접근방법과 통계기반 접근방법으로 구분할 수 있다. 언어학적 접근방법은 어휘사슬(lexical chain)이나 담화트리(*discourse tree*) 등을 이용하여 문서의 담화 구조(discourse structure)를 파악한 다음 요약을 제시하는 방법이다[7, 8]. 통계기반 접근방법은 단어의 빈도, 제목, 문장의 길이, 문장의 위치, 실마리단어나 구(phrase) 등을 특성(feature)으로 사용하여 각 문장이나 문단의 중요도 값을 구하여 그 값이 높은 문장이나 문단을 요약으로 제시하는 방법과 이 두 가지를 혼합한 접근방법이 있다[2, 1]

III 문서 길이 정규화를 이용한 문서 요약 자동화

정보 검색이 문서 집합에서 사용자의 요구에 적합한 문서를 찾아내는 것이라면, 문서요약은 검색된 문장 집합에서 그 문서의 내용을 대표하는 몇 개의 문장을 찾아내는 작업으로 생각할 수 있다.[9].

본 논문에서는 이러한 문서요약을 하는 과정에 질의에 적합한 문장과 문서 길이 정규화를 적용한 문장을 각각 산출하여 신뢰도가 더욱 높은 최적화된 문서 요약을 얻을 수 있음을 보이고자 한다. 제안하는 모델의 개략적인 구성도는 [그림 2]와 같다.



[그림 1] 문서 길이 정규화를 이용한 문서 요약 자동화 시스템의 개요

3.1 벡터 및 초기 질의 생성

사이버 공간의 무한한 정보들 중에서 사용자는 정보 검색 시스템을 이용하여 질의에 대한 문서정보들을 접근할 수 있다. 이러한 문서 집합이 주어지면 문서를 제목과 각 문장들로 분할한 후, 제목과 각 문장을 벡터로 표현한다. 제안하는 시스템에서는 온라인 상의 특정 HTML문서의 소스를 품사 태거를 이용하여 HTML 태그와 명사를 제외한 품사를 제거하고, 제목과 본문의 문장들에서 명사만 추출하여 제목벡터와 문장벡터를 생성한다.

포괄적 요약을 생성하는 경우에는 위에서 생성된 제목벡터를 초기질의(Q^0)로 사용하고, 사용자 주도 요약을 생성하는 경우는 사용자 질의와 제목벡터를 이용하여 초기질의를 구성한다.

3.2 질의 적합 문장 추출

초기 질의와 각 문장 사이의 유사도 계산은 정보 검색에서 많이 사용하는 코사인 유사도(cosine similarity)를 이용한다.

$$CS(S_j, Q^0) = \frac{S_j \cdot Q^0}{|S_j| \times |Q^0|} = \frac{\sum_{i=1}^{t} \omega_{ij} \cdot \omega_{iq}}{\sqrt{\sum_{i=1}^{t} \omega_{ij}^2 \cdot \sum_{i=1}^{t} \omega_{iq}^2}} \quad (5)$$

여기에서, S_j는 각 문장 벡터, Q^0는 초기 질의 벡터를 의미하고, ω_{ij}와 ω_{iq}는 단어 i가 각각 문장과 초기 질의에서 갖는 가중치이다. t는 각 문장 벡터와 초기 질의 벡터를 생성하는데 사용된 단어의 종 개수이다. 식(5)에 의한 유사도 값에 따라 문장을 내림차순 정렬한 후 유사도 값이 큰 상위 k개의 문장을 적합 문장으로 간주한다. 유사도가 같을 경우에는 문서에서 먼저 나오는 문장을 선호한다. 또한, 문서 D_i에 나타나는 단어 T_j(초기 질의 Q^0)에 대한 가중치 요소 ω_{ij}는 다음 식(6)으로 구한다.

$$\omega_{ij} = tf_{ij} \times [\log_2(\frac{N}{df_j}) + 1] \quad (6)$$

여기에서, tf_{ij}는 T_j가 D_i에 나타난 빈도수이고, N은 문서 집합의 문서 개수를 나타내며, df_j는 T_j가 나타나는 문서의 개수이다.

3.3 문서 길이 정규화

식(5)에서 추출한 요약문은 초기 질의에 대한 문장 가중치가 높으면 적합 문장으로 추출되고 또한 요약문이 될 가능성이 높아진다. 그러나 산출된 적합 문장이 원문서에서 주제문이 아닐 수도 있는데, 이를 보완하기 위하여 문서 길이 정규화 방식 중 하나인 최대 단어 빈도 정규화를 도입하여 해결하였다.

$$MN(S_j) = \frac{\sum_{i=1}^{m} CS(tf_m)}{m} \quad (6)$$

tf는 문서에서 초기 질의(Q_0)를 제외하고 출현 빈도가 높은 단어를 나타내고 m은 추출된 tf중 임계치에 해당하는 tf의 개수 중 최초 질의어를 제외한 질의어 수를 나타낸다.

3.4 적합 문서 산출

질의 적합 문장 추출과 최초 질의를 제외한 최대 단어 빈도 정규화를 이용하여 각 문장들의 가중치를 산출하고 이 두 가중치를 바탕으로 질의에 대한 신뢰도가 높은 문서를 선택한다. 즉, 최초 질의에 대한 코사인 유사도는 문서에서 질의를 포함한 문장을 나타내고 여기에 원문서에서 출현이 빈번한 단어 중 질의어를 제외한 단어들의 최대 단어 빈도 정규화 가중치의 평균을 구하여 합산하면 문서에서 요약문을 높은 신뢰도로 산출할 수 있다.

따라서, 적합 문서 산출은 다음 식(7)의 값이 높은 문서를 선택한다.

$$SD(Q^0) = CS(S_j, Q^0) + MN(S_j)$$

$$= \frac{\sum_{i=1}^{t} \omega_{ij} \cdot \omega_{iq}}{\sqrt{\sum_{i=1}^{t} \omega_{ij}^2 \cdot \sum_{i=1}^{t} \omega_{iq}^2}} + \frac{\sum_{i=1}^{m} CS(tf_m)}{m} \quad (7)$$

3.5 요약 추출

최초 질의와 최대 단어 빈도 정규화를 이용하여 문서의 각 문장과의 유사도를 계산하여 요약 문장 후보 리스트를 k개 생성한다.

식(7)에 의한 가중치 값에 따라 가중치가 큰 상위 문장들로 요약 문장 후보 리스트를 각각 생성하는데, 각 요약 문장 후보 리스트 내에 포함되는 문자의 개수는 사용자가 원하는 요약문의 개수와 같게 한다.

사용자가 원하는 요약문이 m개인 경우 각 m개의 후보 문장으로 구성된 문장 후보 리스트에서 요약문을 추출한다. 요약문은 다음과 같은 선호도에 따라 추출한다.

1. k개의 요약 후보 문장 리스트에서 자주 나온 문장
2. 빈도가 같을 경우 문서에서 먼저 나오는 문장

이 선호도에 따라 m개의 문장을 선택하여 최종 요약 문장으로 제시한다.

Ⅳ 실험 및 평가

실험에는 KORDIC의 신문기사 문서집합[1] 중 486건의 문서를 사용하였다. 이 문서집합은 모두 1,000건의 문서로 구성되어 있다고 보고되었으나 이 중에서 500건의 문서는 어절 중간에 줄바꿈 문자가 삽입되어 있는 등 그대로 사용할 수 없는 형태였기 때문에 제외시켰고, 요약이 빠져있거나 내용이 없는 등 올바르지 않은 문서 14건을 제외하여 486건의 문서만 실험에 사용하였다. 이 실험 문서집합의 통계적인 특성은 [표 1]과 같다.

대상 영역	신문기사
문서 개수	486건
문서의 평균 길이	16.46 문장
요약의 평균 길이	4.51 문장
제목의 평균 길이	7.64 개 (명사)
문장의 평균 길이	12.80 개 (명사)

[표 1] KORDIC 문서집합의 통계적인 특성

성능 평가의 척도로는 정확률과 재현율, F-점수를 사용하였다. 올바른 요약문의 총 개수를 S, 시스템이 제시한 요약문의 총 수를 O, 모델이 제시한 올바른 요약문의 개수를 S_0라고 하면, 정확률 P, 재현율 R, F-점수는 다음과 같다.

$$P = \frac{S_0}{O} \qquad (8)$$

$$R = \frac{S_0}{S} \qquad (9)$$

$$F_\beta(R,P) = \frac{(\beta^2+1)PR}{\beta^2 P+R} \qquad (10)$$

β는 정확률과 재현율의 중요도를 조절하는 상수이다.
본 실험에서는 정확률과 재현율의 중요도를 균등하게 하여 F_1-점수를 사용한다. F_1-점수는 다음과 같다.

$$F_1(R,P) = \frac{2PR}{P+R} \qquad (11)$$

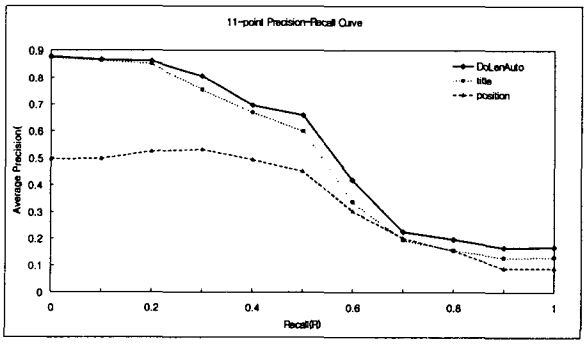
문장벡터가 $S_j=(\omega_{1j},\omega_{2j},\cdots\omega_{nj})$ 같이 구성된다고 하면 각 문장의 가중치는 다음 식(12)로 산출할 수 있다.

$$\omega_{ij} = \frac{freq_{ij}}{\max freq_j} \qquad (12)$$

$freq_{ij}$는 단어 i의 문장 S_j에서의 출현 빈도를 의미하고, $\max freq_j$는 문서 S_j에서 출현한 단어 중 최대 빈도를 의미한다.

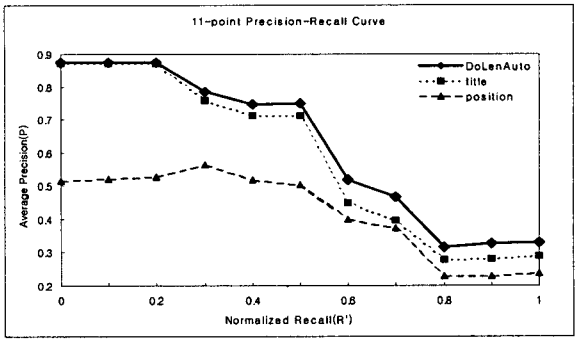
이 방법에 따라 모든 문서의 평균 F_1-점수를 측정한 결과 0.505329를 얻을 수 있었다. 이는 요약문의 정확률과 재현율의 어느 한 부분으로 치우치지 않음을 의미한다.

문서집합들의 각 문장 에 대한 가중치를 식(12)로 산출하여 결정하고 다른 요약 방법들과의 성능을 비교 실험한 결과가 [그림 2]와 [그림 3]과 같다. [그림 2]는 시스템이 제시하는 요약의 길이를 전체 문서 길이의 30%로 하여 정답 요약의 개수와 동일하게 한 경우의 11-포인트 재현율-정확률 곡선이다. 이에 대해 [그림 3]은 요약 길이를 4개의 문장으로 고정하여 실험한 경우의 11-포인트 재현율-정확률 곡선이다.



[그림 2] 30% 요약문에 대한 11-포인트 재현율-정확률 곡선

1) KORDIC에서 소프트과학 프로젝트의 일환으로 수집한 문서요약 테스트 문서집합이다. 이 문서집합은 신문기사로 구성되어 있다 [12].



[그림 3] 요약 길이가 4문장인 경우 11-포인트 재현율-정확률 곡선

기존의 검색 시스템에서 주로 사용하는 방법인 문서의 앞부분을 요약으로 제시하는 방법(position)은 성능이 가장 낮았고, 제안하는 시스템(DoLenAuto)의 성능이 가장 높았다. 제목을 초기질의로 사용하고 초기질의와의 유사도 상위 첫문장에 대해서 문서길이 정규화를 이용한 경우는 초기질의인 제목만 가지고 유사한 문장을 요약으로 제시하는 방법(title)보다 오히려 성능이 저하됨을 알 수 있었다. 이 문서 길이 정규화를 이용하면 질의를 제외한 다른 질의 벡터로 요약문을 산출하여 두 값을 합산하기 때문인데 이는 요약문의 개수를 정의하는 방법으로 완화시킬 수 있다.

V. 결론

본 논문에서는 정보검색을 통해 획득된 문서들을 일차적으로 벡터 및 초기 질의를 이용하여 적합 문장을 산출하고 문서 길이 정규화를 이용하여 신뢰도가 더욱 높은 문서 정보를 요약화한 요약문을 얻을 수 있음을 보였다. 정보검색을 통해 검색된 문서를 요약하여 신뢰도와 유사도가 높은 요약문은 검색 문서의 신뢰도와 적합도가 우수함을 나타낸다. 이로 인해 정보 과재적의 문제를 다소 해결할 수 있으며 최적의 정보를 이용하여 의사 결정에 큰 도움이 된다.

원문서에서 질의어에 대한 코사인 유사도와 문장의 가중치가 높은 문장들과 최초 질의를 제외한 출현 빈도가 높은 단어들이 포함되어 있는 문장을 추출하여 요약문을 산출할 수 있다.

본 논문에서 제안한 문서 요약 자동화 시스템은 단순히 문서의 앞부분만을 보여주는 기존의 정보 검색 시스템의 결과 제시 모듈을 대체하여 효과적인 검색 결과를 제시해줄 수 있다.

그러나 이 시스템은 문장들의 최초 질의 가중치와 문서에서 최초 질의를 제외한 출현 빈도가 높은 단어들이 포함되어 있는 문장들의 가중치를 합하여 두 가중치의 합이 높은 문서를 적합 문서로 산출하고 이 문서에 대한 요약문을 추출하는 방식을 따르고 있다. 이 때 산출된 적합 문서의 요약문이 문서 전체 문장이 될 확률도 고려하지 않을 수 없다. 이는 요약문을 추출하는 과정에서 사용자가 요약문의 개수를 정의하는 방식으로 해결하였다.

참 고 문 헌

[1] Eduard Hovy and Chin-Yew Lin, "Automated Text Summarization in SUMMARIST", In Inderjeet Mani and Mark Maybury, eds, *Advance in Automatic Text Summarization*, pp81-94, The MIT Press, 1999.

[2] Julian Kupiec, Jan Pederson, and Francine Chen, "A Trainable Document Summarizer", In *Processing of ACM-SIGIR'95*, pp.68-73, 1995.

[3] Anastasis Tombros and Mark Sanderson, "Advantages of Query Biased Summaries in Information Retrieval", In *Processings of ACM-SIGIR'98*, pp.2-10, 1998.

[4] Amit Singhal, Chris Buckley, and Mandar Mitra. "Pivoted document length normalization", Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 21-29. Association for Computing Machinery, New York, August 1996.

[5] Gerard Salton. "Automatic text processing-the transformation, analysis and retrieval of information by computer", Addison-Weeley Publishing Co., Reading, MA, 1989.

[6] Gerard Salton and Chris Buckley. "Term-weighting approaches in automatic text retrieval", Information Processing and Management, 24(5):513-523, 1988

[7] Daniel Marcu, "Discourse trees are good indicators of importance in text", Advances in Automatic Text Summarization, pp.123-136, The MIT Press, 1999.

[8] Regina Barzilay and Michael Elhadad, "Using Lexical Chains for Text Summarization", Advances in Automatic Text Summarization, pp.111-121, The MIT Press, 1999.

[9] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, "Summiarizing Text Document : Sentence Selection and Exaluation Metrice", In *Proceeding of ACM-SIGIR'99*, pp.121-128, 1999.