Proceedings of KFIS 2001 Spring Conference, 2001. 5. 19

문서 길이 정규화를 이용한 문서 요약 자동화에 관한 연구

A Study on 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

요 약

WWW(World Wide Web)와 온라인 정보 서비스의 급속한 성장으로 인해, 보다 많은 정보가 온라인으로 이용 혹은 접근 가능해졌다. 이런 정보홍수로 접근 가능한 정보들이 과잉되는 문제가 발생했다. 이러한 과잉 정보 현상으로 인하여 시간적 제약이 뒤따르며 이용 가능한 모든 정보를 근거로 중요한 의사 결정을 내려야 한다. 문서 요약 자동화(Text Summarize Automation)는 이 문제를 처리하는 데 필수적이다. 본 논문에서는 정보 검색을 통해 획득한 문서들을 일차적으로 문서 길이 정규화를 이용하여 질의에 적합하고 신뢰도가 더욱 높은 문서 정보를 얻을 수 있음을 보인다.

Key Words : 정규화, 문서 요약 자동화

I 서론

WWW(World Wild Web)과 온라인 정보 서비스의 급속한 성장으로 인해, 보다 많은 정보가 온라인으로 이용 혹은 접근 가능해졌다. 이로 인해 정보가 과잉되는 현상이 발생하여 원하는 문서의 접근은 많은 시간적 제약을 따르게 한다. 따라서 정보 이용자는 검색 엔진을 통해 문서를 검색하여 정보를 획득할 때 중요한 의사결정을 내려야한다. 문서 요약을 통한 의사 결정은 정보 이용자의 시간적 제약을 해결해 줄 수 있다. 일반적으로 정보 검색시스템들은 문서의 제목과 앞부분만을 보여주어 이 문제를 해결하지만, 이 정도는 정보 사용자가 검색 결과 문서의 적합성을 판단하기에는 미흡하다. 문서 요약 자동화 시스템은 사용자가 원하는 정보를 검색하는데 소요되는 시간을 단축시킴으로써 정보 과재적(information overload) 문제에 대해 효과적인 해결책을 제시해줄 수 있다[1,7].

본 논문은 다음과 같이 구성되어 있다. 2장에서는 문서 요약 자동화와 문서 길이 정규화 분야에 수행되었던 관련 연구들을 고찰하고 3장에서는 문서 길이 정규화를 이용하여 문서 요약 자동화에 대한 모델을 제안하며 4장에서 결론 및 향후연구를 기술한다.

II 관련연구

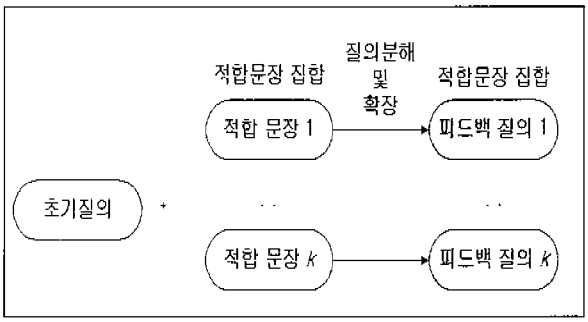
문서 요약이란 문서의 기본적인 내용을 유지하면서 문서의 복잡도, 즉 문서의 길이를 줄이는 작업이다[3]. 요약(Summarization)이란 원 문서에서 중요하다고 여겨지는 것을 선택하고, 혹은 일반화하는 방식으로 내용을 축소, 변형하는 작업이다. 이러한 요약에서 질의를 이용하는 것은 중요하다.

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 문서요약 자동화

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 담화구조(discourse structure)를 파악한 다음 요약을 제시하는 방법이다[6,8]. 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[3,5]

질의에 대한 관점으로 보면 질의 확장과 질의 분해의 방법으로 문서요약을 할 수 있다. 질의확장을 이용한 접근방법은 통계기반의 접근방법이면서도 모델이 난잡해지는 문제를 해소할 수 있는 방법이다. 이 방법은 문서요약을 적합한 문장의 선택 작업으로 간주하여, 정보검색에서 사용하는 질의 확장기법을 문서요약에 적용한 것이다. 적합문장을 이용하여 초기질의를 확장할 때 적합문장 전부를 초기질의에 한꺼번에 적용하지 않고 적합문장 각각을 개별적으로 질의확장을 적용하여 적합문장 개수만큼의 질의로 분해하는 방법도 있다. 질의분해를 이용한 방식은 [그림 1]과 같이 k개의 적합문장들에 대해 각각 개별적으로 질의확장을 수행하여 k개의 피드백 질의를 생성하여 노이즈 문장에 대해 좀더 배타적인 요약을 생성하는 방식이다[2].



[그림 1] 질의분해 및 확장

이 방식은 다음과 같은 수식으로 나타내어진다.

$$Q_i^{new} = Q_0 + S_i \quad (i = 1, \ldots, k)$$

Q_i^{new}는 새로이 확장되는 질의 벡터이고, Q_0는 초기질의 벡터, S_i는 i번째 적합문장 벡터를, k는 적합문장의 개수이다.

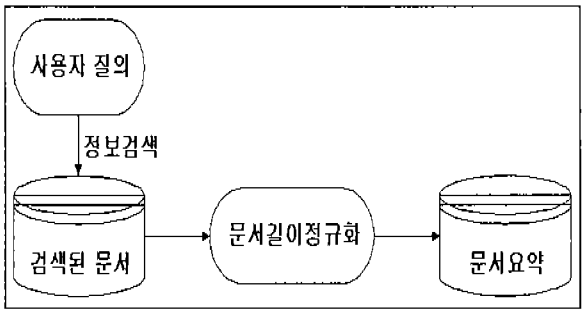
2.2 문서길이정규화

대표적인 정규화 방법에는 벡터 공간 모델에서 가장 일반적으로 사용되는 코사인 정규화(Cosine Normalization)와 SMART 시스템과 INQUERY 시스템에서 단어빈도(tf)를 그 문서에서 가장 많이 출현한 단어의 단어 빈도로 정규화 시키는 최대 단어빈도 정규화(Maximum tf Normalization), 그리고 Okapi 시스템이 TREC에 참가하면서 제안한 바이트 길이 정규화(Byte Length Normalization)가 있다[4].

III 문서 길이 정규화를 이용한 문서 요약 자동화

정보 검색이 문서 집합에서 사용자의 요구에 적합한 문서를 찾아내는 것이라면, 문서요약은 문장 집합에서 그 문서의 내용을 대표하는 몇 개의 문장을 찾아내는 작업이다.

본 논문에서는 이러한 문서요약을 하는 과정에 문서 길이 정규화를 적용하여 최적화된 문서요약을 얻을 수 있음을 보이고자 한다. 제안하는 모델의 개략적인 구성도는 [그림 2]와 같다.



[그림 2] 문서길이정규화를 이용한 문서요약

제안하는 시스템에서 문서 길이 정규화 부분은 코사인 정규화와 최대 단어빈도 정규화를 이용하여 문서요약을 최적화 할 수 있다.

$$C(S_n) = \sqrt{w_1 + w_2 + w_3 + \cdots + w_{n-1} + w_n}$$

$C(S_n)$는 원 문서에서 각 문장이 질의에 대한 신뢰도를 나타내는 코사인 정규화이고 w는 문서의 문장 순서에 따른 가중치이다. $C(S_n)$의 가중치가 높을수록 요약문이 되는 확률은 높아진다. 질의를 이 식에 이용하여 $C(Q)$를 구하면 질의에 해당하는 요약문을 얻을 수 있다. 이 질의에 대한 각 문장의 코사인 정규화 가중치가 높은 문장을 문서 요약문으로 채택을 한다. 그러나 요약문으로 추출된 문장이 원문서에서 주제문이 아닐 수 도 있는데 이

를 보완하기 위하여 최대 단어 빈도 정규화를 도입하여 해결할 수 있다.

$$MN(S_n) = C(Q) + \frac{\sum_{i=1}^{m} C(tf_m)}{m}$$

*tf*는 문서에서 질의를 제외하고 출현빈도가 높은 단어를 나타내고 *m*은 추출된 *tf*중 임계치에 해당되는 *tf*의 개수 중 질의어를 제외한 개수를 나타낸다. 질의를 코사인 정규화 가중치를 이용해서 각 문장들의 가중치와 문서의 최대 단어빈도 정규화를 이용한 문서에서 출현빈도가 높은 단어를 포함하는 문장은 요약문이 될 수 있는 확률이 높아지고 요약문을 최적화 시킨다.

질의에 대한 코사인 정규화 가중치는 문서에서 질의를 포함한 문서를 나타내고 여기에 원 문서에서 출현이 빈번한 단어 중 질의어를 제외한 단어들의 코사인 정규화 가중치들의 평균을 구하여 합산하면 요약문의 신뢰도는 더욱 높아진다.

Ⅳ 결론 및 향후 연구

본 논문에서는 정보검색을 통해 획득된 문서들을 일차적으로 문서 길이 정규화를 이용하여 질의에 적합하고 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.

원 문서에서 질의어에 대한 코사인 정규화 가중치가 높은 문장들 중에서 출현 빈도가 높은 단어들이 포함되어 있는 문장을 추출하여 요약문을 산출할 수 있을 것이다. 그러나 이 방법은 문장들의 두 집합 중 교집합이 되는데 이 교집합이 ϕ가 되면 요약문의 신뢰도는 0%가 될 확률이 높아진다. 이는 요약 표현으로 변환하는 과정에서 중복을 허용하지 않는 모듈을 설계함으로써 해결할 수 있다.

Ⅴ 참고문헌

[1] 한경수, 백대호, 임해창, "질의확장을 이용한 자동 문서요약", 제27회 정보과학회 춘계 학술발표 논문집(B), 제27권, 1호, pp.339-341, 2000.

[2] 한경수, "질의분해를 이용한 적합성 피드백 기반 자동 문서요약", 고려대학교 컴퓨터학과 석사학위논문, 2000.

[3] Julian Kupiec, Jan Pedersen, and Francine Chen, "A Trainable Document Summarizer", Proceedings of ACM-SIGIR'95, pp.68-73, 1995.

[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] 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

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

[7] Anastasios Tombros and Mark Sanderson, "Advantages of Query Biased Summaries in Information Retrieval", Proceedings of ACM-SIGIR'98, pp.2-10, 1998.

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