질의기반 자동문서 요약

김금영*, 강인호***, 안동언*, 정성종*, 박순철**
*전북대학교 컴퓨터공학과
**전북대학교 정보통신공학과
***한국과학기술원 전산학과
e-mail:com94@duan.chonbuk.ac.kr

Query_Based Automatic Text Summarization

Gum-Young Kim*, In-Ho Kang***, Dong-Un An*,
Sung-Jong Chung*, Sun-Cheol Pak**
*Dept of Computer Engineering, Chonbuk-Buk University
**Dept of Infomation and Cummunication,
Chonbuk-Buk University
***Dept of Computer Science, KAIST

요약

웹에 대한 이용이 폭발적으로 증가하면서, 정보검색의 중요성도 증가하고 있다. 이에 따라 정보검색을 효율적이고 신속하게 수행할 수 있도록 다양한 기법이 개발되고 있다. 문서요약은 주어진 문서의 양을 효과적으로 줄이는 기법으로 최근 정보검색 분야에서 활용되고 있다.

본 논문에서는 주어진 질의에 대하여 문서를 요약할 수 있는 자동문서 요약 시스템을 제안한다. 제안하는 시스템은 사용자의 질의에 관련있는 내용만을 포함하는 사용자 주도 요약(user-driven summary) 결과를 산출한다.

1. 서론

인터넷에 대한 이용의 폭발적 증가로 인해 사용자가 원하는 정보를 얻기 위해서는 많은 노력이 필요하다. 현재 많은 검색엔진들이 이러한 노력을 덜어주기 위해 다양한 정보검색 기법을 개발하고 있다.

하지만, 최근 정보검색 분야는 연구와 기술개발의 한계에 부딪쳐 기존의 것과는 다른 새로운 기법이 필요한게 사실이다. 이중 문서분류 분야와 문서 요약 분야는 현재 활발히 연구되고 있는 분야이며, 몇몇 검색엔진에서는 상용화 단계까지 이르고 있다.

문서 요약은 주어진 문서의 기본적인 내용을 유지하면서, 문서의 복잡도를 줄이는 작업이다. 따라서, 검색엔진에서 요약문이 제시된다면, 사용자는 좀더 효율적이고, 편리하게 원하는 정보를 얻을 수 있게 된다.

요약은 기능에 따라 사용자의 적합성 판단에 도움을 주는 지시적 요약(Indicative Summary)과 문서의 중요 내용을 유지하여 그 문서의 대용으로도 사용될 수 있는 정보적 요약(Informative Summary)으로 나눌 수 있다. 또 요약이 제시되는 방법에 따라 문서 내용전체를 포괄하는 포괄적 요약(Generic Summary)과 사용자 질의에 따라 질의에 관련 있는 내용만을 포함하는 사용자 주도 요약(User-driven Summary)으로 나누어 볼 수 있다.

본 논문은 사용자로부터 주어진 질의와 관련있는 내용을 제시하는 사용자 주도 요약을 구현하고 있다. 제시하는 요약문은 단어의 빈도수와 역문서 빈도수를 이용하여 주어진 문서내의 문장을 추출하는 방식이다.

본 논문의 구성은 2장에서 관련연구를 기술하며, 3장에서는 구현한 시스템에 관해 개략적으로 설명하였으며, 4장에서는 논문 문서집합으로 시스템을 실험 및 평가한 결과에 대해 기술할 것이며, 5장에서

는 결론 및 향후연구에 대해 논의한다.

2. 관련연구

자동 문서요약은 1960년대부터 연구되기 시작했으며, 접근 방법에 따라 문장추출 기반 요약 시스템과 정보추출 기반 요약 시스템으로 나눌 수 있다. [1]

문장추출 기반 요약 시스템은 중요 문장을 결정하기 위해 통계적인 방법, 위치나 단서단어를 이용하여 중요문장을 선택하여 요약문을 생성한다.

문장추출 기반 요약 시스템은 통계적 정보를 이용한 요약, 문장특성을 이용한 요약, 의미구조를 이용한 요약, 수사구조를 이용한 요약으로 나눌수 있다.

통계적 정보를 이용한 요약은 문장을 단어 벡터로 표현하고, 단어의 출현 빈도와 말뭉치내의 단어 사용빈도를 이용하여 문장을 추출하는 방식이다.

[2] [3] 문장특성을 이용한 요약은 학습문서로부터 중요 문장을 추출하여 요약문에 나타날 수 있는 어휘 특성과 확률 정보를 이용하여 요약문을 생성한다. [4] 의미구조를 이용한 요약은 WordNet을 이용하여 텍스트 내의 중요 문장과 개념들의 연관관계가 높다는 가정을 이용한다. 수사구조를 이용한 요약은 사람들이 글을 쓸 때, 논리적인 흐름을 가지고 글을 써 나간다는 가정하에 텍스트 내 문장들을 구조화하여 구조화된 표현에서 집중성을 이용하여 요약문을 생성한다. [1]

문장 추출기반 요약 시스템과는 반대로 정보 추출기반 문서 요약 시스템은 문서의 종류에 따라 추출되어야 할 개념들이 정해져 있고, 이러한 개념을 문서내에서 추출하여 주어진 템플릿을 채우는 방식으로 요약문을 생성한다.

질의기반 문서요약 시스템에 대한 연구로는 질의에 대해 의사 적합성 피드백, 제목, 문서의 첫 문장등을 이용하여 질의확장 기법이 있다. [5]

본 논문은 단어빈도수에 기반한 통계적 정보를 이용한 문서 요약 시스템이나, 고빈도 단어가 중심어가 아닐 수도 있기 때문에 영향을 줄이기 위해 추가적인 상수를 사용한다. 또한 본논문은 정보검색에 사용된다는 가정하에 주어진 문서집합내에서 역문서 빈도수를 가중치 계산을 위한 변수로 추가하였다. 또한 웹에서의 사용을 가정하고, 문장의 단위는 임의의 수의 어절수로 하였다.

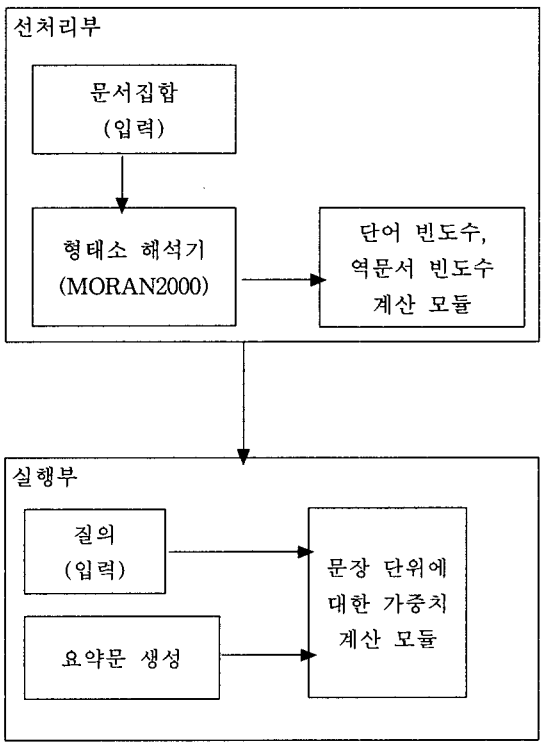
3. 질의기반 자동문서 요약

본 논문에서 제안하는 질의기반 자동문서 요약은 주어진 1개 이상의 명사로 이루어진 질의에 대해서 관련있는 문장단위를 추출하는 작업이다. 관련있는 문장단위를 추출하는 방법은 질의를 이루는 명사를 포함하는 문장단위에 대해서 단어출현빈도와 역 문서 빈도수를 가중치로 계산하여 선택 추출한다.

4. 시스템 구성

질의기반 문서요약 시스템은 선처리부와 실행부로 나누어질 수 있다. 선처리부는 주어진 문서집합을 입력으로 받아 형태소 해석기(MORAN2000) [6]가 문서 집합에서 명사를 추출하는 모듈과 단어 출현빈도수와 역문서 빈도수를 계산하는 모듈로 나누어질 수 있다.

실행부는 입력모듈, 각 문장단위 가중치 계산 모듈, 요약문 생성 모듈로 나눌수 있으며, 선처리부에서 계산되어진 통계값을 참조하여 요약문을 생성한다.



[그림 1] 질의기반 문서요약 시스템의 구성

이때, 각 문장 단위(Passage)에 대한 가중치 계산은 단어 출현빈도수와 역문서 빈도수를 이용한다. 통계적 방법을 요약기법에 이용하는데 있어서 단어의 빈

도수가 가중치에 미치는 영향을 줄이기 위하여 수치를 낮게 조정한다.

$$Tf_{ij} = \frac{tf_{ij}}{tf_{ij}+2.0} \quad (\text{i : 문서, j : 문장})$$

새롭게 계산된 단어빈도수(Tf_{ij})와 역문서 빈도수를 이용하여 가중치를 계산한다.

$$W_{ij} = \frac{\sum_{i=1}^{|passage|}(Tf_{ij} \times idf(w_{ij}))}{|Passage|}$$

Passage 는 각 문장의 길이가 다를때 가중치의 값을 일반화 시키기 위하여 사용되었다. 본 논문에서는 문장 단위(passage) 수를 고정하였으므로 위 식에서 Passage 로 나누는 부분은 사용하지 않는다.

4. 실험 및 평가

실험에 쓰여진 문서집합은 정보과학회 논문 50개를 사용하였다. 이 문서집합은 각기 저자가 직접 작성한 요약을 포함하고 있다. 본 실험에서는 시스템이 제시한 요약문과 저자가 작성한 요약을 비교하였다.

문서를 요약할 때, 고정된 압축률을 사용하는 것보다 고정된 길이를 사용하여 더 나은 성능을 보여주는 연구결과가 있었다. [7] 시스템이 제시한 요약문은 문서의 길이에 상관없이 문장 단위 5개를 추출하였으며 문장 단위안의 어절의 수는 8로 고정하였다. 어절은 형태소 분석기(MORAN2000)이 제시하는 명사를 기본 단위로 하였으므로 동사나 형용사등은 포함하지 않으므로 실제 문서에서의 어절단위와 차이가 있을 수 있다. 또한 제시한 요약문의 가중치가 같을 때는 문서 앞부분에 나오는 문장 단위를 우선적으로 추출하였다.

시스템에 입력되는 질의어는 명사 2-7개로 수동으로 작성한 요약문과 비교하기 위하여 제목에서 중요단어를 뽑았다.

수동으로 작성한 요약문과 시스템이 제시한 요약문의 비교하기 위해서 먼저 두가지 형태의 요약문을 형대소 분석기로 명사만을 추출하였다. 새롭게 처리를 한 요약문을 실험실 연구원 5명이 평가를 하여 A, B, C 세 등급으로 나누었다.

	A	B	C
문서 수	8	19	23

[표 1] 수동요약문과 자동요약문의 비교
(등급 순위 : A>B>C)

■ 논문 제목

지능형 서식 검색 시스템을 위한 자연어 대화 인터페이스

■ 수동요약문

통신수단의 발달로 세계가 정보를 공유해가는 현대사회에서 검색 시스템은 컴퓨터를 전문적으로 사용하는 사용자에게 뿐만 아니라, 일반 사용자에게도 꼭 필요한 시스템이 되었으며, 이들을 위해 사용하기 쉽고, 편리한 인터페이스가 요구되어 지고 있다. 본 연구는 관계형 데이터베이스(relational database)가 구성되어 있지 않은 1000 여 개의 편지문서를 대상으로 자연어 대화 인터페이스와 검색 시스템을 효율적으로 통합한 자연어 대화를 이용한 지능형 서식 검색 시스템에 관한 것이다. 이 시스템은 검색 시스템과 사용자간의 인터페이스를 자연어 대화 인터페이스로 설계하여 사용자에게는 쉽고 편리한 인터페이스를 제공하고, 자연어 대화 인터페이스 시스템의 구성요소를 검색 시스템과 공유하여 사용하도록 함으로서, 시스템을 보다 효율적으로 설계, 구현 하였다. 또 서신 검색 영역에서 발생하는 언어적 지식을 이용한 지능형 검색으로 사용자에게 보다 적절한 결과를 제시 할 수 있게 하였다.

■ 입력 질의어

자연어 + 인터페이스

■ 자동 요약문

<PASSAGE#1>
사용자에게 편리한 자연어 대화 인터페이스를 사용하고, 자연어 대화
<PASSAGE#2>
사용자에게 쉽고, 편리한 인터페이스를 제공하고, 자연어 검색 시스템의 인터페이스와
<PASSAGE#3>
효과적인 의사 소통 수단인 자연어 대화를 이용한 인터페이스이다[6,7,8]. 자연어
<PASSAGE#4>
기존 검색 시스템의 사용자 인터페이스는 사용자가 사용하기에 불편하다. 자연어
<PASSAGE#5>
사용자에게 불편함을 주었다. 자연어 검색 시스템은 사용자의 자연어 질의를

[그림 2] 자동 요약문의 예

본 논문에서 제안하는 질의기반 문서요약은 주어진 질의에 대해 관련있는 문장을 추출하는 것이므로, 작성자가 전체 문서를 포괄적으로 요약하는 수동문

과 비교대상이 되기 어렵다. 하지만, 실험결과의 가시성을 위해 수동으로 요약한 논문과 비교하였다.

5. 결론 및 향후 연구

본 논문에서는 질의에 기반한 자동문서 요약시스템을 제시하였다. 질의에 관련된 문장을 추출하여 요약문을 제시함으로써 문서의 질의와의 관련여부를 쉽게 판단할 수 있도록 하였다.

실험결과 수동 요약문과는 비교대상이 어려웠지만, 질의와 관련있는 문장을 추출한다는 사실을 알 수 있었다. 이 사실을 입증하기 위해 좀더 나은 결과비교방법에 대한 개발이 필요하다.

참고문헌

[1] 이유리, 최기선 "수사구조를 이용한 텍스트 자동요약" 제11회 한글 및 한국어 정보처리 학술대회 1999

[2] 강상배 "한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현" 제9회 한글 및 한국어 정보처리 학술대회 1997

[3] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, Jaime Carbonell "Summarizing Text Documents: Sentence Selection and Evaluation Metrics" In Proceedings of SIGIR, pages 121-128, 1999

[4] 장동현, 맹성현 "문서 구조 정보를 이용한 확률 모델 기반 자동 요약 시스템", 제9회 한글 및 한국어 정보처리 학술대회, pages 15-22, 1997

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

[6] 최유경, 안동언, 정성종 "다중 스레드를 지원하는 한국어 형태소 해석기" 제13회 한글 및 한국어 정보처리 학술대회 2001

[7] Vibhu Mittal, Mark Kantrowitz, Jade Goldstein , Jaime Carbonell "Selecting Text Spans for Document Summaries: Heuristics and Metrics" In Proc. of the 16th National Conference on Artificial Intelligence. pages 467-473 1999