단어의 공기 관계 그래프를 이용한 문서 요약 시스템의 구현

류 제[1] 선복근 박보아 한광록
호서대학교 벤처전문대학원
ryuje@mail.hoseo.ac.kr
bksun@shinbiro.com
winji77@hanmail.net
krhan@office.hoseo.ac.kr

Implementation of summarization system for documents by using a word co-occurrence graph

Je Ryu[1] Bok-Keun Sun Boh-A Park Kwang-Rok Han
Dept. of GSV, Hoseo University

요 약

본 논문은 문서의 내용을 요약하기 위한 시스템의 구현에 대해서 다룬다. 문서의 내용을 분석하기 위해서는 문서의 키워드를 추출하고, 추출된 키워드를 사용하여 문서의 핵심 내용을 찾는 두 가지의 작업이 이루어져야 한다. 본 논문에서는 키워드를 추출하기 위해 형태소 분석 및 전처리기, 그리고 단어의 공기 관계 그래프를 이용한 키워드 추출기를 이용하였으며, 추출된 키워드를 이용하여 문서의 핵심 문장을 찾아내는 핵심 문장 추출기, 그리고 추출된 문장을 분석하여 내용을 요약할 수 있도록 해주는 구문분석기가 이용된다.

1.서론

현재 많이 이용되어지는 인터넷 서비스 중에는 문서의 검색 및 분류 서비스가 상당수를 차지하고 있다. 그러나, 이러한 인터넷 사이트들에서 검색결과로서 제공되어지는 문서에 대한 설명이 부적절하거나 미약하며, 이로 인해 사용자가 원하는 문서를 찾기 위해서는 검색결과로 제공되어지는 문서를 모두 확인해야하는 어려움이 있다. 이에 이러한 웹 사이트에서 제공되어지는 문서의 내용을 간략하게 요약하여 검색 결과로서 함께 제공되어지는 것은 필수불가결한 요소가 되었다[7].

문서의 내용을 요약하기 위해서는 우선 문서의 내용 중에서 키워드를 추출해야 한다. 본 논문에서는 키워드를 좀더 효과적으로 추출하기 위하여 단어간의 공기관계를 이용한 방식을 키워드 추출 방법에 적용하였다.

키워드를 추출하고 나면 추출된 키워드를 포함하는 문장 혹은 문장의 핵심이 되는 문장을 추출하여야 한다. 일반적으로 문서 내에서 키워드를 포함하는 문장이 문서의 핵심내용일 확률이 높으며, 문장에 많은 수의 키워드를 포함할수록 그 확률은 더욱 높아진다. 이렇게 추출된 핵심 문장들을 구문분석기를 이용하여 불필요한 요소를 제거 혹은 문장의 내용을 정리하여 문서의 내용을 요약하게 된다.

2.관련연구

문서의 내용을 요약하는데 있어서 그 결과의 정확성을 좌우하는 것은 정확한 키워드의 추출에 달려 있다. 또한, 본 논문에서의 키워드는 문서의 내용을 요약하는데 있어서 문서의 주제 혹은 저자의 주장을 대표할 수 있어야 한다.

근래에 가장 일반적으로 사용되어지는 방법으로는 통계정보에 의한 키워드 추출법과 같이 문서상에서의 단어의 출현 빈도를 이용하여 키워드를 추출하는 방법이 있다. 그러나 이러한 단어의 출현빈도에 의존한 계산은 문서분류에는 유용하게 쓰일 수 있으나[5], 본 논문에서 구현하는 문서의 내용을 요약하기 위한 키워드 추출에는 부적합하다. 이는 단어의 출현빈도가 문장의 독자적인 주장을 나타내는 단어간의 상관관계를 표현하는 것은 아니기 때문이다. 대부분의 기존의 키워드 추출방법은 문서의 내용을 요약하기에는 부적합하기에 본 논문에서는 단어와 단어간의 상관관계를 계산하여 문서를 분석하고 주제어 혹은 저자의 주장에 해당하는 단어와 밀접한 관계를 찾아내어 키워드를 추출하는 단어의 공기관계 그래프를 이용한다[4].

3. 공기관계 그래프

본 논문에서는 저자의 주장을 나타내는 키워드 자동 추출 방법으로서 공기관계 그래프를 이용한다. 이 방식은 고속 알고리즘은 아니지만 문서가 저자의 독자적인 생각을 주장하기 위해 쓰여졌다는 가정하에 문서상의 저자의 주장을 대표하는 키워드 추출에 효과적인 방식이다.

3.1 공기 관계 그래프의 구성

공기관계 그래프의 구조는 문서를 크게 개념클러스터

(Mean-Cluster), 주장(Insistence), 전개(Deployment)로 구분한다[3].

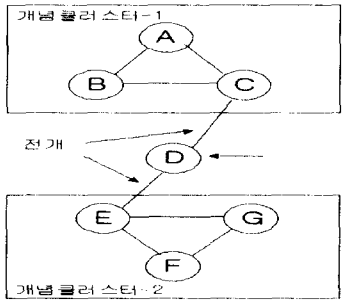

그림 1. 공기관계 그래프의 구성

개념클러스터는 문서가 기초로 하고 있는 기본 개념으로서 문서의 형성을 준비 또는 전제로 한다. 예를 들어, 출현 빈도가 높은 단어들은 문서의 요점과는 상관관계가 적을 수는 있으나 문장을 작성하는데 당연시되는 전제들이다. 주장은 저자가 의도하는 문서의 핵심이 될 수 있으며, 개념클러스터와 강하게 연결되어 문장을 통합하는 역할을 가진다. 전개는 개념클러스터와 주장을 이어주는 역할을 하며 내용의 중요한 흐름을 나타낸다.

3.2 키워드 추출

공기관계 그래프는 개념클러스터와 주장을 이어주는 전개의 강도가 높은 단어를 키워드로 선택한다. 문서를 준비 또는 전제로 하는 기본 개념클러스터에 속하는 단어들 중에서 저자의 의견을 나타내는 주장과 관련이 가장 높은 단어가 문서의 키워드로 선택되는 것이다. 다시 말하면, 주장과 강하게 연결되어 있는 개념클러스터 내의 단어가 키워드가 된다.

4. 문서 요약 시스템의 구현

본 논문에서 구현된 문서요약 시스템은 크게 웹 에이전트, 형태소 분석 및 전처리기, 키워드 추출기, 핵심문장 추출기, 구문분석기의 5개 모듈로 구성된다.

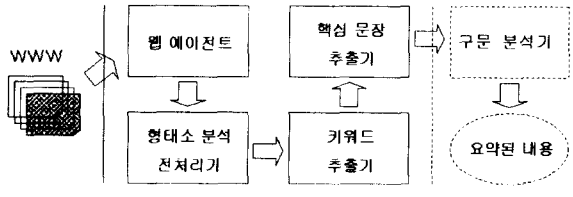

그림 2. 문서요약 시스템 구성도

우선 웹 에이전트는 웹 상의 문서를 수집한다[8]. 본 시스템에 사용된 웹 에이전트는 자체에 HTML Tag 등을 제거하는 기능을 가지고 있기 때문에 형태소 분석기에서의 Tag제거 작업을 생략할 수 있다. 일단 수집된 문서들은 형태소 분석 및 전처리기를 통하여 키워드 추출의 기초작업인 후보단어를 분류한다. 본 시스템에서는 각각의 품사중에서도 명사, 형용사, 동사만을 키워드의 후보단어로 설정하며, 단어들에 대해서 중복을 제거하고, 기타 불필요한 요소를 제거한다[1]. 이렇게 각각의 준비 모듈을 통해 선정된 후보단어에 대해서 키워드 추출기는 단어간의 공기 관계 그래프를 생성하고, 이를 이용하여 키워드를 추출한다. 추출된 키워드는 핵심 문장 추출기에서 해당 키워드들을 이용하여 핵심문장을 추출하게 되며, 이렇게 추출된 핵심 문장들은 구문분석기를 통하여 정리되어져 문서의 내용을 요약하게 된다[2](그림2에서 점선으로 표시된 부분으로서 현재 구현중이다).

4.1 준비단계

대상 문서의 수집은 웹 에이전트를 통해서 이루어 진다. 수집된 문서는 키워드를 추출하기 위해서 우선 불용어(Stop List), HTML의 Tag등 불필요한 단어를 제거하며, 중복을 제거한다. 본 논문에서는 숙어의 요소는 고려하지 않았으며, 단어의 선정도 명사, 형용사, 동사로 제한하였다. 후보단어의 집합은 다음과 같이 표현한다.

$$\text{Document} = \{W_1, W_2, W_3, W_4, \ldots\ldots\ldots W_n\}$$

4.2 키워드 추출

1) 개념클러스터(Mean-Cluster)의 형성

개념클러스터를 형성하기 위해서 우선 각 단어들로 이루어진 그래프를 생성한다. 그래프는 문서내에서 출현빈도가 높은 상위 30 개의 단어를 추출하여 그래프의 Vertex로 간주한다. 추출된 Vertex들의 공기관계를 나타내기 위해서 Vertex들의 쌍을 만들어 공기도를 계산한다. 계산된 공기도가 유효한 경우에만 Vertex간의 관계를 나타내는 Edge를 생성해준다. 공기도는 두 개의 단어가 문서내의 모든 문장에 대해 동시에 출현한 횟수의 합을 의미한다.

$$Co(W_i,W_j) = \sum_{s\in d} |W_i|_s |W_j|_s$$

$|X|_s$는 문장 S에서의 요소 X의 출현 빈도

일단 그래프를 생성하면 그래프 상의 단어들의 연결관계를 분석하여 개념클러스터를 생성한다. 개념클러스터는 그래프 상에서 루프를 형성하는 단어들의 집합으로 표현된다[6].

2) 주장(Insistence)의 형성

주장을 계산하기 위한 함수 Key(W)는 임의의 단어 W가 각각의 개념클러스터에 속하는 단어들에 대해서 동시에 출현할 수 있는 확률을 계산한다. 본 논문에서는 Key(W)값이 높은 상위 12개의 단어를 주장으로서 선택하였다.

$$Key(W) = [\ 1 - \prod_{c}^{Clusters} (1 - f(W,C)/F(C))\ ]$$

$$f(W,C) = \sum_{s \subset D} |W|_s |C-W|_s \,, \quad F(C) = \sum_{C \subset D} \sum_{W \subset S} |C-W|_s$$

$$|C-W|_S = |C|_s - |W|_s \quad \text{if} \quad W \subset C$$
$$= |C|_s \quad \text{if} \quad W \not\subset C$$

Bases : 개념 클러스터의 개수
f(W,C) : 단어 W와 개념클러스터 C와의 공기도
F(C) : 모든 단어와 모든 개념클러스터의 공기도의 합.
S : 문장 W : 단어
$|C|_s$: 개념클러스터 C에 포함되는 단어의 문장 S에서의 출현빈도.

3) 키워드 추출

키워드의 추출은 그래프를 이용해서 얻어진 주장중의 단어 W_i와 개념에 포함되는 단어 W_j 사이의 강도를 계산하여 강도가 높은 단어를 키워드로 선택한다. 본 논문에서는 단어 12개를 키워드로 선택하였다.

주장과 개념간의 강도(W_i, W_j) :

$$Co(W_1,W_1) = \sum_{s \subset d} |W_i|_s |W_j|_s$$

KeyWord = min(12,|Document|)

4.3 핵심문장의 추출

추출된 키워드들은 핵심문장 추출기의 후보 문장을 선정하는 기준이 된다. 본 논문에서는 동일한 개념 클러스터내에 속하는 키워드들 중에서 두 개 이상의 키워드가 동시에 포함되는 문장을 핵심문장으로 하였다.

5. 실험

아래의 표들은 테스트 수행 결과 중 시사영역에 있는 문서를 분석한 결과이다.

[표 1] 표본 문서(총 38라인, 총 단어수 296개)

이번 주 박씨는 다시 뉴스의 초점을 받게 된다. 박바지에 계획이 변경되지 않는다면, 다부진 이 한국인은 하원 윤리 위원회의 공개 회의에서 증언을 하게 된다.
...(중략)....
그의 증언은 1년 6개월 전부터 드러나기 시작한 한국의 뇌물 스캔들을 최초로 완벽하게 미국 국민들에게 보여주게 된다.

[표 2] 추출된 키워드

받다, 되다, 박, 뇌물, 하다, 위원회, 공개, 미국, 한국, 하원, 증언, 특별

[표 3] 추출된 핵심문장

박바지에 계획이 변경되지 않는다면, 다부진 이 한국인은 하원 윤리 위원회의 공개 회의에서 증언을 하게 된다. 윤리 위원회의 특별 고문이며 워터게이트 사건으로 유명한 레온 죠워스키의 조사를 받게 되면, 박은 31명의 의원 명단을 공개할 것으로 기대된다. 박은 이들이 한국에 대해 지속적으로 미국이 군사·경제 원조를 지지해주고 또한 국제적인 쌀 중개인으로서 그의 개인적인 위치를 지지해 준다는 댓가로 75만 달러의 뇌물을 받았다고 주장한다.

6. 결론 및 향후과제

본 논문에서는 문서 요약 시스템의 구현에 대해서 다루었다. 앞의 실험결과는 그 결과가 비교적 우수하게 나온 것을 보여준다. 앞에서도 언급했듯이 본 논문에서 구현한 시스템의 경우, 그 결과의 정확도는 키워드 추출의 정확도에 많은 영향을 받는다. 본 논문에서 이용된 키워드 추출기는 키워드의 후보로서 기존의 키워드 추출기와는 달리 명사 외에도 동사, 형용사 등을 키워드로 함께 추출하기 때문에 그 정확도를 기존의 키워드 추출방법과 비교 평가하기 어려운 점이 있으나, 문서의 내용을 요약하기에는 기존의 키워드 추출방법을 사용하는 것보다는 그 정확도가 우수한 것으로 평가된다.

본 논문에서 구현된 문서 요약 시스템은 현재 핵심 문장을 추출하는 단계까지 구현되었다. 추출된 핵심 문장 내에서 불필요한 요소를 제거하고, 한 개 이상의 문장을 유연하게 연결해 줄 수 있는 구문 분석기를 이용한 완전한 문서 요약 시스템은 현재 구현중이다.

7. 참고문헌

[1] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 대학원 컴퓨터공학과 박사 학위 논문, 1993.

[2] 서영훈, 이하규 외, "한국어 구문 Tagged Corpus 구축 및 구문 분석 데이터 사전 개발", 한국 전자 통신 연구소 최종 연구 보고서, 1998.

[3] 류 제, 한광록 "단어의 공기 관계 그래프를 이용한 인터넷 문서의 키워드 추출", HCI2000 학술대회발표논문집 9권 1호, pp894-899, 2000.

[4] Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida "Automatic Indexing by Segmentation and Unifing Co-occurrence Graphs" 전자정보통신학회논문지 D-1 vol. J82-D-I, No.2, pp391-400, 1992

[5] 조광제, 김준택, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동분류", 정보과학회 봄 학술발표논문집 4권 2호, pp508-510, 1997

[6] Ellis Horowitz, Sartaj Sahni, Dinesh Mehta "Fundamental of data structures", p330-396

[7] 강현규, 박세영, "정보검색", 한국정보처리학회 p37-47, 1998.9

[8] 이성민, 김태윤 "A news on demand service system based on robot agent" Proceedings of the 1998 international conference, p.528-532, 1998