계층적 개념 트리를 이용한 문서 분할 기법

이병희[O] 최익규* 박승규* 김민구**
아주대학교 정보통신 전문대학원[O]*, 아주대학교 정보 및 컴퓨터 공학부**
{acecult[O], ikchoi, sparky}*@ajou.ac.kr, minkoo@ajou.ac.kr**

Text segmentation using concept hierarchy tree

Byonghui Lee[O] , Ikkyu Choi, Seungkyu Park
Graduate School of Information and Communication, Ajou University,
Minkoo Kim
College of Information & Computer Engineering, Ajou University

요 약

문서 분할 기법은 문서 내에 존재하는 다양한 주제들을 자동적으로 추출하는 기법이다. 이 분야의 연구는 크게 사전적 관계에 근거한 기법과 통계적 데이터에 근거한 기법으로 나누어져 연구되어 왔다. 사전적 관계에 의한 기법은 단[1]어들의 사전적 의미와 관계에 근거한 기법이고, 통계적 데이터에 의한 기법은 주로 단어들의 분포를 이용한 기법이다. 여기에는 몇가지 문제점이 있는데 사전적 관계에 근거한 경우에는 분산된 주제들을 통합하여 추출하기 어렵고, 통계적 데이터에 근거한 기법은 정확한 주제의 개수를 찾기 어렵다는 점이다. 본 논문에서는 계층적 개념 트리를 이용하여 보다 정확한 개수의 주제들을 찾아낼 수 있는 문서 분할 기법에 대해 소개하고자 한다.

1. 서 론

문서는 대개 여러가지의 주제로 구성되어 있다. 이러한 예는 신문기사 모음이나, 웹 문서, 잡지의 기사들 등에서 쉽게 찾을 수 있다. 문서내에 존재하는 주제를 찾아내는 것은 정보 검색이나 문서 자동 요약 기술에 응용될 수 있는 기술이다 [5]. 정보 검색에 있어서 사용자들은 대개 검색된 결과 문서 전체보다는 자신이 원하는 내용만을 검색하기를 원한다. 이러한 요구를 만족시키기 위해서 문서는 관련 주제에 따라 내용이 분할되어야 한다. 문서 자동 요약에 있어서 문서의 요약은 핵심 주제에 관련된 문장 및 단어들의 집합이라고 볼 수 있다. 핵심 주제에 관련된 문장 및 단어를 추출하기 위해서는 우선 관련 주제에 따라서 내용을 분할 할 수 있어야 한다. 따라서 문서의 내용을 관련 주제에 따라 분할 하는 것은 문서 자동 요약을 위한 기초 기술에 해당한다. 보다 정확하게 문서의 내용을 분할 할 수 있다면 정보 검색 및 문서 자동 요약 기술에서 큰 도움이 될 수 있다 [4].

본 논문의 구성은 2장에서 관련 연구에 대한 분석을 살펴보고, 3장에서 기본적인 문서 분할 기법과 계층적 개념 트리를 이용하는 방법에 대해 설명한다. 4장에서 실험 결과 및 평가 방법을 살펴본 후 5장에서 결론 및 향후 과제에 대해 설명한다.

2. 관련 연구

문서 분할 기법은 크게 두가지 방법으로 구분할 수 있다. 첫째는 사전적 결합도를 이용한 방법으로 주로 시소러스를 이용하여 개념적으로 관계가 있는 단어들을 클러스터링하여 문서 분할에 이용한다. 둘째는 통계적 데이터를 이용하는 방법으로 어떤 주제의 처음과 끝의 위치를 찾아내기 위해 단어의 분포도 혹은 단어들의 쌍의 분포 같은 통계적 데이터를 이용한다.

2.1 사전적 관계 기반 연구

사전적 결합도란 단어들 간의 사전적, 문법적 관계의 정도를 의미한다. 사전적 결합도에 기반한 많은 연구들은 거의 한가지 정도의 형식을 이용한다. TextTiling[2]이 그러한 연구 중 하나이다. TextTiling은 문서를 고정된 크기의 블록으로 분할한 후에 각각의 유사도를 계산한다. 또한 각 블록간의 유사도를 이용하여 관련이 높은 블록끼리 연결하여 문서 분할을 수행한다. 유사도를 계산하기 위해 시소러스를 이용하여 단어들 간의 관계를 구한다. 또한 Ponte와 Croft[1]는 이웃한 문장들 간에 함께 나타나는 단어들의 빈도를 이용하여 문서 분할을 수행한다. 대개의 문장에는 단어의 수가 그리 많지 않기 때문에 이 연구에서는 LCA(Local Context Analysis)라는 기술을 이용하여

본 논문은 KISTEP의 국가지정연구실 사업의 일환으로 지원받아 수행되었음. (과제번호 M10302000087-03J0000-04400)

문장을 확장한 후에 함께 나타나는 단어의 빈도를 구하고 있다. 문장 간에 함께 나타나는 단어의 빈도를 이용하여 문장 간의 유사도를 계산하고 이 유사도 값을 이용하여 각 주제를 구분 지을 수 있게 된다.

2.2 통계적 데이터 기반 연구

통계적 데이터에 기반한 연구들에서는 명사나 구 등의 통계적 정보를 이용한다. 많은 연구들이 단어의 출현 빈도를 이용하여 중요 단어를 가려내고, 몇 개의 단어로 이루어진 단어 구(phrase)를 찾아내어 분할에 이용한다. 단어 구는 하나의 단어가 가지는 모호성을 줄여주기 때문에 이용가치가 있다 [3]. 단어의 출현 빈도 및 분포 등의 정보를 주로 이용하기 때문에 문서의 내용적 속성에 대해서 유연한 성능을 보여준다. 사전적 관계에 기반한 연구들의 경우에는 대상 문서의 내용을 고려하여 시소러스를 선택해야 하지만 통계적 데이터 기반의 경우에는 시소러스는 거의 이용하지 않기 때문에 융통성 있는 성능을 가질 수 있다. 통계적 데이터 기반의 연구에서 가정하고 있는 것은 서로 다른 주제는 서로 다른 단어의 분포를 보인다는 것이다 [2]. 이러한 가정에 근거하여 어떠한 영역을 대표할 수 있는 단어의 분포, 단어 쌍의 빈도, 고유 명사의 위치 등의 정보를 이용하여 문서 분할을 수행한다 [3].

2.3 분석

사전적 관계에 기반한 문서 분할 기법은 우선 구현하기에 쉬운 장점이 있다. 그러나 주로 선형 탐색에 의존하다 보니 분산된 주제들에 대해서는 분할이 어려운 문제가 있다. 통계적 데이터 기반의 분할 기법은 분산된 주제들에 대한 분할은 비교적 성공적으로 수행하지만 특정 단어 및 단어 쌍의 분포를 이용하기 때문에 정확한 구분은 쉽지 않다. 따라서 비교적 정확하게 분할 작업을 수행할 수 있으면서 분산된 내용들 또한 연결시켜서 분할 할 수 있는 방법이 필요하다. 3장에서 이러한 방법에 대해 논해보고자 한다.

3. 계층적 개념 트리를 이용한 문서 분할 기법

2장에서 거론한 문제점을 해결하기 위해 본 논문에서는 두 단계의 작업을 수행하였다. 첫째는 사전적 관계를 이용한 선형 탐색에 의한 분할 작업이고, 둘째는 선형 탐색에 의해 분할된 블록들에 대해 계층적 개념트리를 이용하여 분산되어 있는 블록들을 관련 주제에 따라 연결시켜 주는 작업이다.

3.1 시소러스를 이용한 선형 분할

두 단계 중 첫번째 단계인 선형 분할은 시소러스를 이용한다. 문장의 확장을 위해 시소러스를 이용하게 되는데 빈약한 시소러스를 사용하게 될 경우 좋은 성능을 기대하기 어렵다. 본 논문에서는 프린스턴 대학에서 개발해 현재 광범위하게 사용되고 있는 WordNet 시소러스를 이용하였다. WordNet 시소러스를 이용해 각 문장에 있는 불용어를 제외한 단어들에 대해 다의어, 유의어, 동의어 등으로 확장한 후 각 문장간에 함께 나타나는 단어의 빈도를 이용하여 문장간 유사도를 계산하였다.

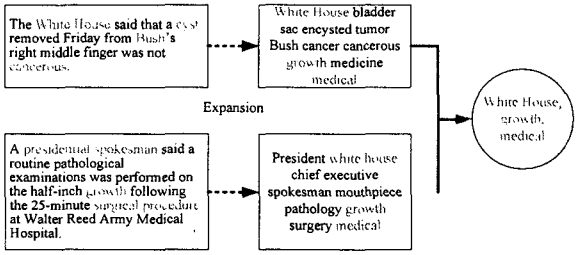

그림 1 시소러스를 이용한 문장간 유사도 계산

이때 문장간 유사도의 임계치를 설정하여 임계치를 밑도는 유사도를 갖는 문장들을 새로운 주제의 첫 문장으로 판단하였다. 이 실험에서 임계치는 단어 빈도수 3으로 설정하였다. 이러한 방식으로 시소러스를 이용한 선형 탐색이 가능하였다.

3.2 계층적 개념 트리를 이용한 문서 분할

선형 탐색을 통하여 구분된 블록들에 대해 각 블록이 어느 정도의 유사도를 갖는지를 계산하기 위해 계층적 개념 트리를 이용하였다. 계층적 개념 트리는 야후의 카테고리를 이용하여 구성되었다. 야후의 카테고리는 도메인 전문가들에 의해 설계되었기 때문에 합리적으로 구성되어 있다. 우선 야후의 카테고리를 이용하여 트리를 구성하고 각 카테고리의 대표 페이지들을 수집하여 색인 작업을 수행하였다. 야후에는 너무나 많은 카테고리와 페이지들이 있기 때문에 본 연구에서는 5 단계까지의 카테고리에 속하는 페이지만을 수집하여 색인하였다. 5 단계까지의 색인 결과 61,000 여개의 컨셉 노드들이 구성되었다 [6].

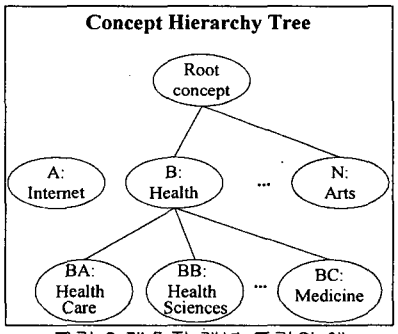

그림 2 계층적 개념 트리의 예

이러한 트리와 정보검색의 유사도 계산 기법을 이용하여 선형 분할의 결과로 추출된 블록들이 어느 노드에 가장 가까운가를 계산 할 수 있게 되어 문서 내에 분산되어 있는 같은 주제를 다르고 있는 블록들을 찾아낼 수 있게 된다. 각 블록이 속하는 노드 간의 유사도는 트리 내에서의 거리를 계산하여 구하는데 이용된 식은 다음과 같다.

$$CD\_WT(n,n') = \frac{MAX\_PATH - SHORTEST\_PATH(n,n')}{MAX\_PATH}$$

MAX_PATH는 현재 구축된 개념 트리 내에서 계산 가능한 최대의 거리이고, SHORTEST_PATH는 두 노드 간에 계산 가능한 가장 짧은 거리를 의미한다. 위와 같은 식을 이용해 계산하면 0에서 1사이의 값을 얻을 수 있으며, 얻어진 값은 클수록 노드 간 유사도가 높다는 것을 의미하게 된다. 이러한 식을 이용하여 어떠한 블록들이 유사한 주제를 다루고 있는 가를 계산할 수 있다.

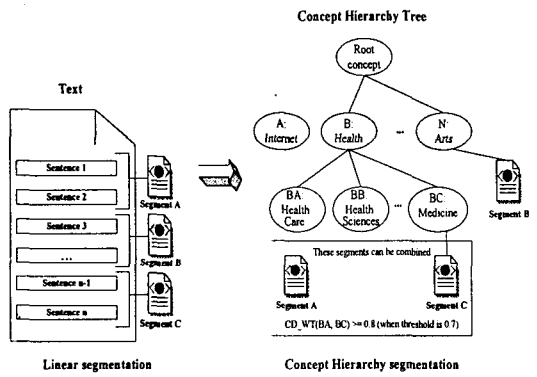

그림 3 계층적 개념 트리를 이용한 문서 분할 시스템

4. 실험 및 결과

본 연구의 실험을 위한 테스트 데이터는 NIST에서 주관하는 TREC 컨퍼런스의 데이터 중 TREC8 문서집합을 이용하였다. TREC8 문서 집합 중 LA 타임즈 기사 모음을 대상으로 하여 실험을 하였다. 기사를 6가지 주제로 분류하여 총 55개의 기사를 하나의 문서로 구성한 후 실험에 이용하였다. 연예, 도시, 재정, 스포츠 등의 주제로 구성된 기사들은 각각 내용이 잘 구분되고 또한 같은 주제를 다루는 기사들도 고르게 분포하기 때문에 실험을 위한 좋은 데이터라고 할 수 있다.

테스트 데이터에 대해 3장에서 설명하였듯이 두 단계로 문서 분할 작업을 수행하였고, 결과의 평가를 위해 정보검색에서 성능평가를 위해 사용되는 Recall과 Precision을 이용하였다. 실험 결과는 다음과 같다.

	원본 기사	추출된 기사
총 기사	55	49 (부분 매치 6)
Precision	0.8775 (43/49)	
Recall	0.8909 (49/55)	

위 테이블에서 부분 매치가 의미하는 것은 정확하게 원본 기사와 일치하게 추출되지는 않았으나 그 내용을 표현함에는 무리가 없을 정도로 추출된 기사를 의미한다.

이와 같이 추출된 기사들에 대해 계층적 개념트리를 이용해 기사들간의 유사도를 구하고, 그 유사도가 임계치(본 실험에서는 0.7로 설정)보다 높게 계산된 경우 같은 주제를 다루는 기사로 판단하여 추출한 결과는 다음과 같다.

기사 번호	야후로부터 추출된 개념
10, 17, 36	Actors and Actress
2, 48	Comedy
4, 8, 24	Health care
15, 32, 44	Mental Health
19, 45	Entertainment
23, 27, 49	Sports
28, 40	Countries
29, 34	Software
31, 42	Business Management

5. 결론 및 향후 과제

본 연구에서는 앞서 언급한 기존 관련 연구들의 문제점을 개선할 수 있는 기법을 소개하였다. 사전적 관계에 기반한 선형 탐색으로는 찾을 수 없는 분산된 주제들에 대해서 추출할 수 있는 기법이 계층적 개념 트리를 이용한 방법이다. 선형 탐색을 통해 비교적 정확한 개수의 주제들을 추출하고 계층적 개념 트리를 이용함으로써 비슷한 주제를 다루고 있는 블록들을 연결하여 문서 내에 분산되어 있는 주제에 대한 추출이 가능하였다. 그러나 역시 정확한 개수의 주제를 찾아내지는 못하였으며, 이는 향후 연구의 과제로 삼을 것이다. 또한 야후의 일부 카테고리를 이용하여 개념 트리를 구성하였는데 차후에는 보다 많은 카테고리와 그 내용을 수집하여 계층적 개념 트리를 구성할 계획이며, 이는 보다 정확한 문서 분할을 가능하게 할 수 있을 것으로 보인다.

6. 참고 문헌

[1] Jay M Ponte and W. Bruce Croft: The segmentation by topic. European Conference on Digital libraries, 1997

[2] Nicola Stoke, *Joe Carthy and Alan F. Smeaton*: Segmenting Broadcast News Stream using lexical chains, Proceedings of Starting AI researchers Symposium, 2002

[3] Jeffery C. Reynar, Microsoft Corporation: Statistical Models for Topic Segmentation, Meeting of the Association for computational linguistics, 1994

[4] Masao Utiyama and Hitoshi Isahara: A statistical Model for Domain-independent Text Segmentation, Meeting of the Association for Computational Linguistics, 2000

[5] Marti A. Hearst and Xerox Palo Alto Research Center: Multi-Paragraph segmentation of expository text, 32nd. Annual Meeting of the Association for Computational Linguistics, 1994

[6] Ikkyu Choi and Minkoo Kim: Topic distillation using *Concept Hierarchy Tree*, SIGIR, 2003