21. 도합 유사도를 이용한 한국어 추출요약 시스템

컴퓨터공학과 김 준 홍
지도교수 김 재 훈

월드와이드 웹(World Wide Wed, WWW)을 중심으로 인터넷의 급속한 팽창과 컴퓨터 보급의 증대 그리고 온라인 서비스의 증가로 인하여 이용 가능한 정보가 폭발적으로 증가하게 되었다. 이러한 호나경에서 정보 검색엔진은 사용자들에게 필요한 정보를 찾아주는 유용한 도구이다. 그러나 일반적으로 검색엔진은 사용자들에게 너무 너무 많은 문서들을 찾아준다. 일례로, 인터넷 환경에서의 웹 검색 엔진들은 사용자의 질의어 하나 당 수십 개에서 수만 개까지의 문서를 찾아 주기도 하는데, 사용자들이 이 모든 문서들을 직접 읽어서 관련성을 판단한다는 것은 매우 힘든 일이다. 오늘날은 너무나 많은 정보로 인하여 사용자가 원하는 정보를 찾기 어렵게 되었으며, 이런 현상을 정보 과적재(inforrmation overload)라고 한다. 이러한 정보 검색 환경에서 효과적인 정보 획득의 수단으로서 자동 요약 시스템에 대한연구가 증가하고 있다.

문서요약은 원문서의 의미를 유지하면서 원문서의 길이나 정보의 복잡도를 줄이는 작업이다. 즉, 문서요약은 정보압축(information compression)이다. 문서요약은 일상적인 생활에서도 널리 사용되고‘있는 방법이다. 예를 들면, 헤드라인 뉴스, 각종 회의의 의사록, 책이나 CD등의 논평 등이 일상적인 생활 속의 문서요약에 대한 예이다.

본 논문은 한국어 문서를 대상으로 하여 의미적으로 중요하다고 판단되는 문장을 추출하여 요약문서를 생성하는 하국어 추출요약 시스템에 대한 사항을 기술하였다. 본 논문에서 문서는 문서관계도라고 하는 그래프로 표현된다. 노드는 문서의 구성 요소인 문장을 뜻하며, 링크는 노드들 간의 의미적 관계를 나타낸다. 의미적 관계는 도합유사도란 다른 노드들간의 유사도의 합을 의미한다. 본 논문에서는 도합유사도를 이용한 한국어 문서 요약 시스템을 제안한다.

본 논문에서는 시스템의 성능을 평가하기 위해서 논문(서론과 결론), 논문(전체본문), 신문 기사의 세 종류의 말뭉치를 사용하였다. 시스템이 생성한 요약문의 크기가 본문 크기의 20%이고, 본문이 논문(서론과 결론)일 경우, 재현율과 정확률은 각각 46.8%와 76.9%를 보였으며, 본문이 논문(전체 본문)일경우, 재현율과 정확률이 83.3%와 19.7를 보였다. 또한 본문이 신문기사일 경우, 재현율과 정확률은 각각 30.5%와 42.3%를 보였다. 또한 제안한 방법은 상용시스템인 워드프로세서에 내장된 문서 요약도구보다 좋은 성능을 보였다.

본 논문에서 제안한 요약 시스템은 단순한 모델을 사용하고 있으며, 구현이 용이하고 쉽게실용적으로로 사용할 수 있다는 장점을 가지고 있다. 본 논문에서 제안한 문서 요약 기술을 정보검색의 색인이나 문서 분류 등의 분야에 응용함으로써 더욱 향상된 검색 시스템을 구현할 수 있을 것으로 기대된다.