제한된 한글 입력환경을 위한 음소기반 근사 문자열 검색 시스템 구현

Implementation of Phoneme-based Approximate String Searching System for Restricted Korean Character Input Environment

윤태진
Yoon Taijin
부산대학교 컴퓨터공학과
ytj@pusan.ac.kr

ABSTRACT

모바일 기기가 발전함에 따라 입력 수단에 대한 연구는 중요한 이슈이다. 키패드, 쿼티키패드, 터치, 음성인식 등 다양한 입력장치가 사용되고 있으나 아직 데스크탑 입력장치에 비해 편의성이 떨어져서 입력시에 오타나 탈자 등의 오류가 포함되는 경우가 많다. 이러한 입력 오류는 문자 메시지 등 사람과의 의사소통에는 문제를 일으키지 않으나 사전, 주소록 등의 데이터 베이스 검색에는 치명적인 오류로서 원하는 검색 결과를 얻지 못하게 된다. 특히 한글의 경우 자음과 모음의 조합을 통해 글자를 생성하는 특성상 1만자가 넘는 글자의 조합이 가능하여 영문에 비하여 오류의 빈도가 높다. 기존의 검색 시스템은 suffix tree등을 이용하여 입력 오류를 처리하지만 다양한 오류에 대응하기에는 한계가 있다. 본 논문에서는 오자, 탈자 등의 입력 오류를 허용하면서 빠른 검색이 가능한 근사 한글 단어 검색 시스템을 제안하고자 한다. 이 시스템은 기존의 알파벳에 적용된 Approximate String Searching을 한글에 효과적으로 적용할 수 있는 여러가지 알고리즘과 기법이 포함되어 있다. 그리고 제안된 시스템을 이용한 변형 욕설 필터링 시스템의 개발에 대해 이야기하고자 한다. 이 시스템은 유저의 각종 변형 욕설 입력에 대해 90% 이상의 필터링 성능을 보였다.

KEYWORDS Hangeul string, approximate string matching, global alignment

1 연구의 필요성

누구나 한 두 글자의 오타로 인해 원하는 검색결과를 얻지 못하는 경험을 가지고 있을 것이다. 특히 모바일 기기의 경우 키패드나 터치 스크린과 같은 제한되고 불편한 입력 수단을 사용하게 되므로 데스크 톱의 경우보다 오타를 입력하게 될 확률이 높다. 안그래도 불편한 입력 장치 때문에 어렵게 입력한 검색어가 오타로 인해 다시 입력해야 할 경우의 스트레스는 이루 말할 수 없을 정도이다. 운전 중에 급하게 네비게이션을 조작하다가 오타로 인해 잘못된 결과가 출력되었을때의 스트레스는 모두 공감할 것이다. 오타가 단순한 조작 미스로 인한 것이 아니라 잘못된 기억으로 인해 발생했을 경우 문제는 더욱 심각해진다. 사회가 복잡해짐에 따라 한 개인이 관리해야 하는 데이터량은 점점 증가하고 있다. 정렬과 검색 시스템의 도움으로 효율적인 데이터의 관리가 이루어지고 있으나 검색에 필요한 keyword는 어디까지나 사람의 기억에 의존할 수 밖에 없다. 그러나 사람의 기억은 불완전하여 때때로 검색에 필요한 정확한 keyword를 도출해내지 못하는 경우가 발생한다. 특히 한글의 경우