웹 사이트 콘텐츠 분석 시스템

Analysis System of Web Contents

백선욱, 성민영, 안성혜
상명대학교

Seonuck Paek, Sung Min-Young, Ahn Sung-Hye
Sangmyung University

요약

사회가 복잡해짐에 따라 처리하고 분석해야 할 콘텐츠의 양은 점점 더 많아지며, 이러한 많은 정보들을 자동적으로 체계적으로 분류하여 필요한 통계를 바로 추출하는 기능이 점점 더 중요해지고 있다. 본 논문에서는 직업 정보 사이트인 jobkorea 사이트에서 IT 분야의 구인 관련 공고를 추출하고 추출된 문서들에 있는 단어들의 통계 처리를 통해 현재 IT 관련 산업체에서 필요로 하는 기술 분야 및 직종을 자동으로 분석하여 보여줄 수 있는 시스템을 개발하였다. 개발된 시스템은 IT 관련 학과의 교과과정 개편 등의 다양한 응용에 활용할 수 있을 것이라 기대된다. 또한, 본 시스템은 직업 정보 사이트의 분석 이외에 콘텐츠 동향 분석이나 관련 분야의 통계 처리를 필요로 하는 다른 사이트에도 쉽게 확대 적용될 수 있을 것으로 기대된다.

Abstract

As the amount of web contents in the Internet increases, it becomes hard to find out statistics that users want. In this paper we propose an analysis system based on the statistics of words, which can be used to prospect trends in a specific area. We applied this system to job recruiting site and we can find out trend and statistics about what part of technology is needed in job market and the result of this paper can be used for an application such as restructuring curriculum of a department in universities. It can also be used to predict trend in other areas.

Ⅰ. 서론

1. 연구 배경

2,000년 말, 노동부에서 집계한 우리나라의 직업 명칭 수는 12,306개에 이르고 있으며, 산업구조와 사회제도의 변화에 따라 직업세계도 끊임없이 변화하고 있다. 이에 따라 대학에서 직업교육훈련, 자격검정, 진로지도 등 인적자원개발의 기초 자료를 확보하기 위해서는 직업세계의 구체적인 변화 양상을 분석해야 하는 필요성이 높아지고 있다.

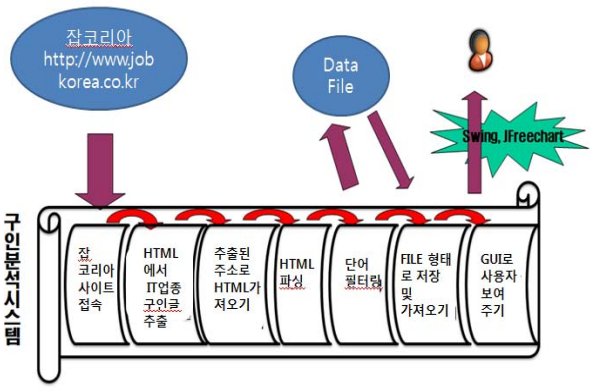
직업세계는 직업의 생성과 소멸, 분화와 통합이 지속적으로 이루어지고 있는 동적인 특성이 있다. 특히 IT 분야의 직종들은 요구되는 기술과 직종의 유동성이 매우 큰 분야이다. IT 분야의 직종들의 동향과 추세를 알고 싶을 땐 jobkorea[3]나 '사람인'[8] 등의 직업 정보 사이트의 글들을 살펴보거나 신문기사 등을 활용할 수 있지만, 이러한 매체를 통해서는 단편적인 정보 만을 얻을 수 있을 뿐이다. 예를 들어 IT 분야에서 현재 가장 필요로 하는 기술이 어떤 것인지를 바로 알기는 쉽지 않다. [5]에서는 사용자의 클릭 수를 각 웹 페이지의 가중치에 누적함으로써 다수 사용자의 검색 행위에 의한 묵시적 평가가 웹 페이지의 검색 순위에 반영되는 검색 시스템을 구현하였다. 이렇게 사용자가 검색을 함으로써 그 단어가 누적되어 통계를 내는 통계 시스템[4]은 많이 있지만 게시된 글들에 나타난 단어들을 대상으로 통계를 내어 분석하는 시스템은 아직 찾기 힘든 실정이다. 한편 [1][2]에서는 웹 사이트의 사용자 이용 현황 분석 등의 분석을 통해 유용한 정보를 얻고자 하였다.

본 논문에서는 jobkorea 사이트를 대상으로 그곳에 등록된 IT 분야의 글들을 추출하여 단어들을 검색하고 통계처리를 하여, 현재 IT 산업체에서 필요로 하는 기술들이 어떤 것들인지 또는 어떠한 동향을 보이고 있는지를 파악할 수 있도록 하는 시스템을 설계 구현하였다.

본 논문의 II절에서는 구현된 직업 정보 사이트 분석 시스템의 주요 기능 및 구조에 대해 기술하고, III절에서는 시스템의 동작 원리에 대해 기술한다. IV절에서는 구현된 시스템의 몇 가지 테스트 결과에 대해 논하고 V절에서는 향후 연구방향에 대해 논하였다.

Ⅱ. 시스템 기능 및 구조

본 시스템은 먼저, '잡코리아' 사이트의 IT 분야의 구인 관련 게시물들을 자동으로 추출한다. 다음으로 추출된 게시물들에 대해 다양한 질의를 통해 통계 결과를 월별/연도별로 그래프 형태로 파악할 수 있도록 해 준다. 이를 통해 시간에 따른 구인 직종 및 필요 기술의 흐름을 파악 할 수 있도록 하였다.



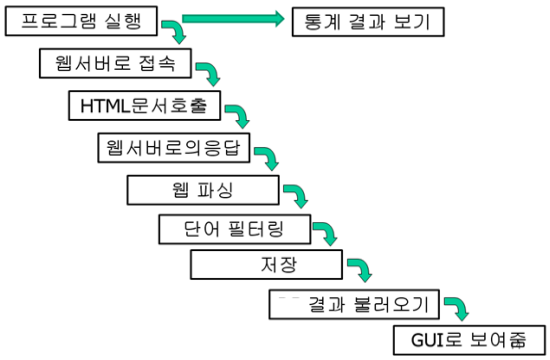
▶▶ 그림 1. 구인 분석 시스템의 구조

본 논문에서 개발된 시스템은 그림 1에 나타난 바와 같이 클라이언트(구인분석시스템)와, 분석 대상인 jobkroea 사이트[3], 처리 결과를 저장하는 데이터베이스 등으로 구성된다.

클라이언트는 JAVA로 구현되었는데, jobkorea 사이트에 접속하고, 구인관련 게시물들의 URL을 자동으로 상위 단계에서 하위단계로 추적한 후에 필요한 게시물들을 추출하도록 한다. 추출된 게시물들은 HTML 태그의 파싱을 통해서 텍스트로 분류되고 분류된 텍스트는 단어 필터링을 통하여 파일에 저장된다. 파일에 저장된 단어들은 통계처리를 거친 후에 JfreeChart[6]를 통해 사용자가 알기 쉽게 GUI형태의 차트로 보여주도록 하였다.

Ⅲ. 시스템 동작

본 시스템은 클라이언트가 사이트에 접속하여 게시물을 가져오는 기능과 가져온 게시물들을 파싱하고 필터링하는 기능, 그리고 필터링된 단어들을 데이터베이스에 저장하는 기능 등으로 나누어진다(그림 2 참조).



▶▶ 그림 2. 구인 분석 시스템의 동작 원리

먼저, 클라이언트 프로그램을 수행하면 '잡코리아' 사이트의 URL로 접속하여 업종/직종별 보기와 연결된 URL의 HTML문서를 가져온다. 가져온 HTML 문서에서 다시 하위 단계로의 URL을 검색하여 단계 별로 HTML 문서를 불러오게 되어 각 카테고리 별 최하위단계의 HTML문서들을 불러와 카테고리별, 지역별로 파싱하게 된다.

다만 최하위단계의 HTML문서는 회사의 인사고과의 개인 정보 및 회사 정보가 쉽게 노출 되기 때문에 보안정책이 설정되어 있다. 때문에 Referer 검사나 서버 스크립트에서 HTML 헤더의 User-Agent를 검사 등을 통해 보안을 피하여 가져오는 방법 또한 필요하다.

'잡 코리아' 역시 위와 같은 웹 보안이 걸려있기

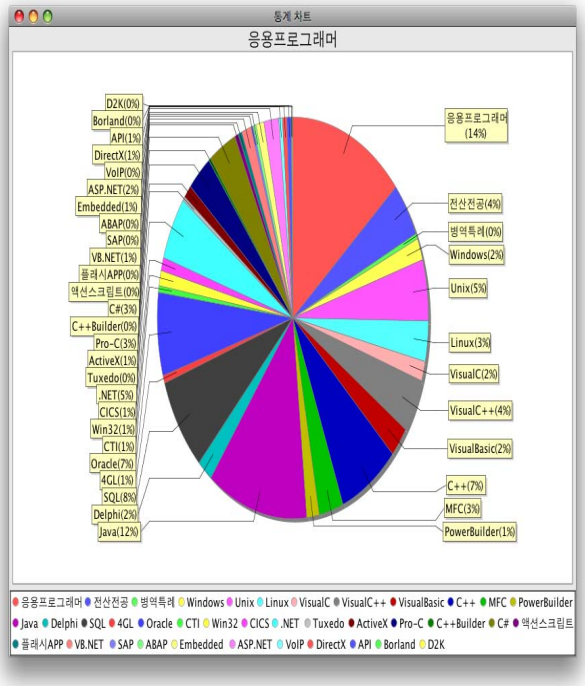
때문에 파이어폭스에서 FireBug나 Live HTTP headers 플러그인을 설치, 헤더의 송수신 내역을 분석해서 보안을 피하고 접근을 할 수 있다.

다음으로 파싱된 단어들은 HTML태그가 제거된 텍스트가 된다. 파싱된 텍스트들은 다시 지역별 카운팅을 통하여 각 단어별로 분류되어 저장된다. 파일에 저장된 단어들은 내부적인 구조화된 형태로 읽어 통계 결과 값들을 받아 JFreeChart로 차트화시켜 다양한 결과 값들을 사용자에게 보여주게 된다.

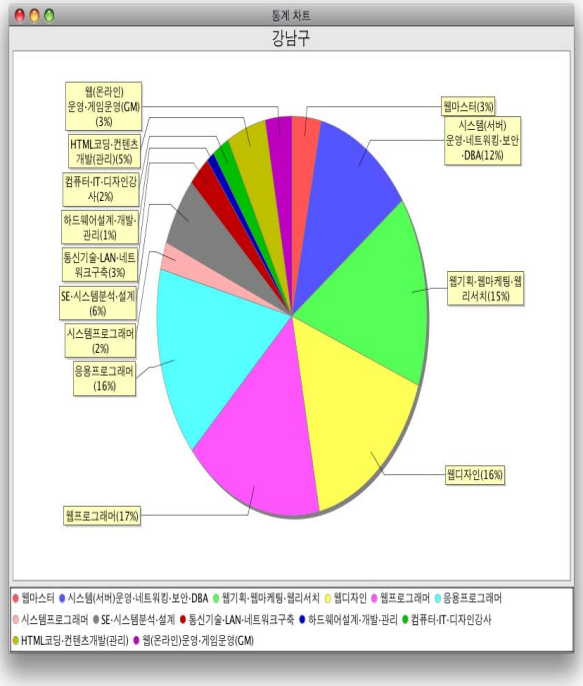
Ⅳ. 테스트

본 절에서는 본 시스템을 사용하여 실험한 몇 가지 결과를 기술하는데, 본 시스템은 직종별 통계, 지역별 통계 및 두 개의 혼합 통계 기능을 지원하다. 본 논문에서는 서울특별시만을 대상으로 한 결과만을 보여주고 있는데, 그 이유는 IT업계의 구인모집이 80%이상이 수도권 그것도 서울에 집중되어 있기 때문이다. 그림 3은 직종별 통계를 보여주고 있으며 그림 4는 지역별 통계를 보여주고 있다. 그림 5는 이 두 가지를 합한 통계를 보여주고 있다. 이와 같은 방식으로 나머지 질의에 대해서도 유용한 통계 결과를 알 수 있다.

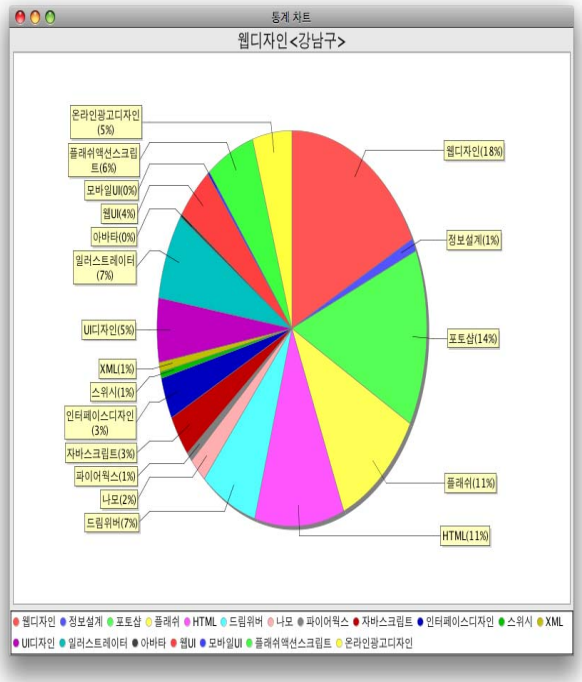
그림 3에서 검색 당시의 구인 게시 글에서 1위 웹프로그래머, 2위가 웹디자인, 3위가 응용프로그래머 임을 알 수 있다. 또한 그림 4에서는 지역별 수요 현황을 알 수 있는데, 검색 당시의 1위는 응용프로그래머이고 2위는 Java, 3위는 SQL프로그래머임을 알 수 있다. 또한, 이 두 가지를 합한 지역군별 직종모집 현황을 그림 5에서 보여주고 있다. 그림 3~5에서 알 수 있듯이 특정 웹 사이트에서 IT 분야의 구인 글들을 추출하여 다양한 질의에 대한 통계 결과 등을 파악할 수 있다.



▶▶ 그림 3. 직종별 통계



▶▶ 그림 4. 지역별 통계



▶▶ 그림 5. 직종 & 지역 통합 통계

V. 결론 및 향후 연구 방향

본 논문에서는 직업 정보 사이트에 접속하여 현재 산업체에서 필요로 하는 기술 및 직종에 관한 통계를 쉽게 파악할 수 있도록 한 시스템을 구현하였다. 본 시스템은 IT 분야의 산업체의 수요 흐름을 한 눈에 파악할 수 있게 하여 졸업생 및 취업준비생, 대학생들에게 구인 정보를 제공해주고 IT 관련학과의 학과 교과과정 설계 시에도 참고할 수 있다. 본 시스템을 활용하면, 구직을 원하는 사용자들이 IT 산업체의 추세를 알 수 있게 하여 취업 준비에 대한 계획 및 준비를 할 수 있게 하는 장점이 있다. 또한, 본 시스템은 IT 분야 외의 다른 직업 분야에도 쉽게 확장 적용할 수 있다.

본 논문에서 개발된 시스템은 현재는 해당 사이트의 정보를 가져올 때에 사이트의 구조에 의존적으로 설계되어있으나, 검색 기능을 강화하여 사이트가 개편되더라도 스스로 찾아는 혹은 관리자를 통해 바뀐 부분을 관리자모드로 직접입력으로 사이트가 개편되더라도 그때그때 적용을 바로 할 수 있게 개선을 할 계획이다. 또한 현재는 통계를 내야 할 단어를 사용자가 지정해주는 방식을 택하고 있는데, 향후에는 관리자의 단어 지정 없이 자동으로 문서를 분석하여 그 문서 안에 들어있는 단어들의 분류 체계를 자동으로 만들고 그에 따른 통계를 내는 방식으로 확장할 계획에 있다. [7]에서 제시하고 있는 주제어 추출 기술과 결합 하거나 보다 나은 알고리즘으로 단어들을 추출 할 수 있다면 최적의 시스템으로 향상될 것이다. 향후에 [8] 등의 사이트와 연계하여 통합 통계를 낸다면 보다 폭넓은 결과를 확인할 수 있을 것이다. 또한, 제안된 시스템을 다른 분야에도 적용할 수 있을 것으로 기대할 수 있다. 예를 들어, 인터넷뉴스 게시판을 대상으로 단어 통계를 낸다면 그 기사에 대한 사람들의 관심도를 쉽게 파악 할 수 있는데, 이러한 관심 분야 사이트에 본 시스템을 적용한다면 사용자에게 다양한 정보들을 알기 쉽게 분류하여 실시간으로 전달할 수 있을 것으로 기대된다.

■ 참 고 문 헌 ■

[1] Bartell, A.L "Using content analysis and Web design heuristics to evaluate informational Web sites: an exploratory study," Professional Communication Conference, Page(s):771 - 777, July 2005.

[2] Zhou, Q. and DeSantis, R. "Usability issues in city tourism Web site design: a content analysis," Professional Communication Conference, Page(s):789 - 796, July 2005.

[3] http://www.jobkorea.com.

[4] 김형일, 김준태, "질의어 의미별 사용자 선호도를 이용한 웹 검색의 성능 향상," 한국정보과학회 논문지 B - 소프트웨어 및 응용 VOL.31 NO.08, 2004.

[5] http://help.naver.com/faq/faqRealtime Keyword.jsp

[6] JfreeChart 〈http://www.jfree.org/〉

[7] 이창범, 김민수, 이기호, 이귀상, 박혁로, "주성분 분석을 이용한 문서 주제어 추출," 한국정보과학회 논문지 B - 소프트웨어 및 응용 VOL.29 NO.10, 2002.

[8] http://www.saramin.com.