군집화를 이용한 하이브리드 기반 채용검색 랭킹 기법

조보연
고려대학교 컴퓨터정보통신대학원 디지털정보공학과
e-mail : cbyangel@gmail.com

Recruiting Ranking Techniques Based on Hybrid Using Clustering

Bo-Yun Cho
Dept. of Digital Information Engineering, Korea University
Graduate school of Computer Information Communication

요 약

인터넷의 활용이 보편화 됨에 따라 정보의 양은 급격히 늘어나고 있다. 이에 취업을 희망하는 구직자의 경우 IR 로부터 원하는 정보를 검색하기 위해 과거보다 더 많은 시간과 노력이 필요하게 되었다. 이에 본 논문에서는 TF(Term Frequency)기법을 통해 문서를 추출하고 추출된 문서의 Doc_ID 빈도수를 기준으로 한 내용기반과 군집기법을 혼합한 하이브리드 검색 시스템을 제안한다. 구직자들이 클릭한 취업정보들의 링크번호들을 K-means 알고리즘을 이용하여 군집화를 한다. 생성된 군집들은 각기 하나의 문서로 가정하고, 기존 문서과 더불어 검색 주제와 연관성을 갖고 있는 문서들을 동적비율로 검색 랭킹 하는 방식이다. 기존의 IR 기술과의 비교 실험을 통해 성능을 평가하였다. 실험결과 본 논문에서 제안한 방법이 기존의 방법보다 우수함을 확인할 수 있었다

1. 서론

인터넷이 발달하고 보편화 됨에 따라 국가기관, 일반기업 및 학교에서는 홍보, 마케팅 및 의견수렴 등의 수단으로 이메일 마케팅 기법을 사용하고 있다. 회원들에게 지속적으로 원하는 정보를 이메일을 통해 정기적으로 보냄으로써 고객들의 만족도를 높이고 고객들의 반응을 수집함으로써 쌍방향 커뮤니케이션이 가능하고 더 나아가 좀 더 나은 다양한 서비스를 제공하기 위한 노력을 계속하고 있다. 이러한 일환으로 사용자들이 검색했던 질의어 로그, 웹 주소등 사용자의 행동 정보를 통하여 생성된 사용자 프로파일(profile)을 가지고 자동으로 질의어를 추천하는 연구가 진행되고 있다[1][2]. 그러나 사용자 프로파일을 가지고 개인화된 검색 결과를 제공 하는 경우, 회원 행동 정보의 양에 따라 실제 검색 결과에 영향을 주지 못하는 문제가 발생하거나 추천된 질의어도 불분명한 의미인 경우도 많다[3][4][5]. 이에 본 연구에서는 내용기반의 하이브리드 검색 랭킹 기법을 제안한다. 메일발송에 대한 회원들의 오픈/링크클릭에 대한 정보들을 수집한 후 구인/구직사이트에서 주로 사용하는 카테고리의 수만큼 몇 개의 그룹으로 나누어 군집화 기법을 수행한다. 즉 유사한 클릭 형태를 가진 그룹들을 자동으로 검출할 수 있다. 검색 시 질의어에 매칭도가 높은 군집과 문서 ID 빈도에 따른 내용기반 랭킹을 혼합하여 채용 검색 랭킹 시스템에 적용 방안을 제시하고자 한다. 본 연구의 2 절에서는 군집과 더불어 취업사이트에서 제공하는 채용정보의 URL 구조적 특성 및 TF 기법을 기반으로 문서를 추출하는 지표에 관한 연구를 소개한다. 3 절에서는 2 절의 연구를 이용하여 하이브리드 기반의 랭킹 기법에 대한 제안 방법을 설명하고, 4 절에서는 제안방법에 관한 실험 및 결과 대해서 기술하고, 5 절에서는 결론 및 향후 연구과제에 대하여 언급한다.

2. 관련연구

2.1 TF 기반 문서 추출

문서는 종종 각 속성이 특정 용어(단어)의 발생 빈도를 표현하는 벡터로 표현된다. 문서들이 수천 또는 수만의 속성(용어)을 가지기는 하지만, 각 문서는 0 이 아닌 속성을 상대적으로 적게 가지고 있으므로 희소해진다. 그 이유는 임의의 두 문서는 다수의 동일 단어를 포함하지 않을 것이고 그러므로 0-0 매치를 세면 대부분의 문서들이 대부분의 다른 문서들과 높은 유사도를 가지게 되기 때문이다. 그러므로 문서에 대한 유사도 척도는 자카드 척도처럼 0-0 매치를 무시할 필요가 있으나, 비-이진 벡터도 처리할 수 있어야 한다. 문서 a 가 w_{a1}, w_{a2}, $w_{a3...}$ 의 키워드를 가지고 있고, 문서 b 가 w_{b1}, w_{b2}, w_{b3} 의 키워드를 가지고 있을 때, a 는 벡터 (w_{a1}, w_{a2}, w_{a3}...), b 는 벡터 (w_{b1}, w_{b2}, w_{b3}...) 로 나타낼 수 있다. 이 때, 문서 a b 간의 코사인 유사도는 식(1)와 같다 [6]

$$\mathrm{COS}(\theta) = \frac{a * b}{\|a\|\|b\|}$$

$$\mathrm{COS}(\theta) = \frac{(Wa1 * Wb1 * Wa2 * Wb2 * Wa3 * Wb3 \dots)}{\sqrt{(Wa1^2 + Wa2^2 + Wa3^2 \dots)} * \sqrt{(Wb1^2 + Wb2^2 + Wb3^2 \dots)}} \quad (1)$$

2.2 채용정보의 URL 구조를 이용한 문서 추출

URL 은 웹 브라우저로 사이트를 열람할 때 주소창에 입력하는 문자열이다. 즉, 리소스의 위치를 지정하는 통일된 기술방법이다. URL 은 크게 스킴(Scheme), 호스트명, 경로 세 부분으로 나눌 수 있다. 스킴은 리소스를 취득하기 위한 방법을 나타내며 RFC1738 이라는 문서에 규정되어 있다. 호스트명은 리소스가 존재하는 호스트의 이름이다. 통상적인 URL 에 '?(물음표)'가 이어져 있다. 이 부분을 쿼리 문자열(Query String)이라고 하며 웹서버에 전달된다. 쿼리 문자열 속은 다시 &(앰퍼샌드)로 나뉘어 있으며 각 부분은 '매개변수명=값'의 형식으로 표현된다.[7] 예를 들어, http://www.wikibook.co.kr?id=113982 이라 하면 해당된 링크의 문서의 id 는 113982 이다. 사용자가 원하는 회사의 로고나 채용정보을 통해 클릭할 경우 각기 다른 링크 ID 를 부여하지만 보여주는 문서의 Doc_ID 는 동일하다. 그러므로 링크 ID 기준이 아닌 문서 Doc_ID 의 빈도수로 랭킹하여 웹 브라우저를 통해 보여준다.

2.3 군집기반 추출

비계층적 클러스터링 기법인 K-means 알고리즘은 K 개의 초기 중심점들을 선택하며, 여기서 K 는 사용자가 명시하는 매개변수로서 원하는 군집들의 개수를 나타낸다. 각각의 점은 가장 가까운 중심점에 지정되며, 각 중심점에 할당된 점들의 집합이 군집이 된다. 그런 다음, 각각의 군집 중심점은 군집에 할당된 점들을 기반으로 하여 갱신된다. 이러한 할당과 갱신 단계를 반복하여 어떤 점도 군집이 바뀌지 않거나 또는 중심점들이 동일하게 유지될 때까지 계속한다[8]. K-means 알고리즘을 수행할 때 유사도 계수인 자카드 계수를 메소드로 사용한다. 채용정보에 관하여 회원들이 읽었는지에 대해 0 과 1 의 이진 속성만을 포함하기 때문이다. 값 1 은 두 객체가 완전히 유사함을, 값 0 은 객체들이 전혀 유사하지 않음을 나타낸다. X 와 y 를 n 개의 이진 속성을 가지는 두 객체라 가정하면 두 이진 벡터의 비교로 다음의 4 개 수(빈도)를 만들어낸다.

f_{00} = x 가 0 이고 y 가 0 인 속성 수
f_{01} = x 가 0 이고 y 가 1 인 속성 수
f_{10} = x 가 1 이고 y 가 0 인 속성 수
f_{11} = x 가 1 이고 y 가 1 인 속성 수

자카드 계수는 비대칭 이진 속성으로 구성된 객체들을 처리하기 위해 사용되며 식(2)과 같다.

$$J = \frac{\text{매칭 존재 수}}{\text{00 매치에 관련되지 않은 속성의 수}} = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} \quad (2)$$

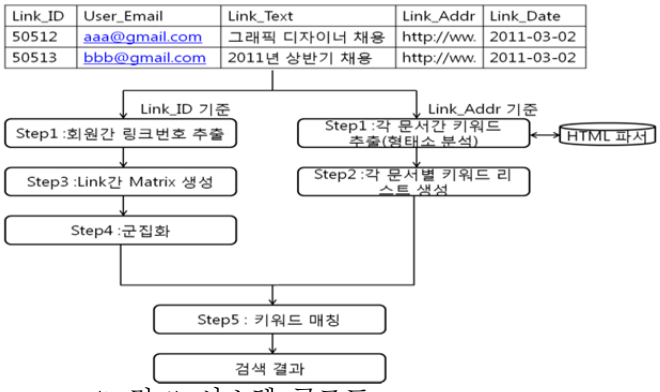
추출된 각각의 군집들의 연관성을 찾기 위해 TF-IDF 기법을 이용하여 대표키워드를 추출한다. TF-IDF 기법은 문서 내에서 질의 단어의 발생횟수(TF)와 문서 집합 내에서 해당 질의 단어가 얼마나 많은 수의 문서 내에서 발견되는지(DF)의 역수(IDF)를 이용한 것으로 기본적인 계산 기법은 식(3)와 같다[9].

$$\text{Score}(d,q) = \sum_{t \in q} \text{TF}(d,t) * \text{IDF}(t) \quad (3)$$

여기서 q 는 질의를 의미하며 t 는 질의 내에 포함된 개별 단어를, d 는 개별 문서를 의미한다.

3. 제안하는 하이브리드 검색 랭킹 기법

본 논문에서 제안하는 하이브리드 검색 랭킹 기법은 (그림 1)과 같이 총 5 개의 step 으로 구성된다.



(그림 1) 시스템 구조도

Step1 인 경우, 회원번호는 회원들의 고유번호이고 링크번호는 메일링 서비스 시 링크된 번호이다. 링크내용은 2 가지로 분류되는데 링크된 텍스트이거나 URL 주소이다. URL 주소일 경우 해당 업체의 로고이거나 상세정보를 보여줄 수 있는 페이지로 이동하고 수집된 정보는 아래 <표 1>와 같이 DB 에 저장된다.

<표 1 > CareerDB 구축 예

User_ID	Link_ID	Link_Text	Link_Addr
150	24014	각 분야별 신입/경력 채용	http://www.career.co.kr?id=113982&l_idx=29774

클릭된 Link_Addr 를 토대로 크롤러(Crawler)를 이용하여 HTML 문서로 저장한다. 저장된 HTML 문서를 HTML parser 를 이용하여 태그의 특정패턴을 분석하여 제거하고 상세요강과 관련된 텍스트만 파일의 형태로 저장한다[10]. 저장된 텍스트로부터 TF 기법을 이용하여 키워드와 키워드의 빈도수를 추출한다. 키워드 추출 시 한국어 분석 모듈 KLT version2.0 을 이용한다[11]. Step2 인 경우, 각 문서 별 추출된 키워드로부터 TF 기법을 이용하여 키워드 벡터에서 키워드의 빈도를 찾는다. 검색을 빠르게 하기 위해 Doc_TermList 에 저장되며 <표 2>와 같다.

<표 2> Doc_TermList 구축 예

Doc_ID	Link_ID	Term	TermCnt
113982	24014	신입	3

Step3 인 Link 간 Matrix 생성인 경우, 회원이 클릭한 링크번호일 경우 1 로 표시하고, 클릭하지 않은 경우 0 으로 한다. 아래 <표 3>에 나타나 있다. 링크 별 벡터를 만들어서 각 링크간의 Similarity 을 계산한 후, Link-Link Similarity Matrix 를 만든다. 예를 들어 Link1 = (1, 0, 0, 1), Link2 = (1, 1, 1, 1) Link1 과 Link2 의

Similarity 는 2 이며 <표 4>와 같다.

<표 3> Link-User Matrix

	User1	User2	User3	User4	....
Link1	1	0	0	1	
Link2	1	1	1	1	
Link3	0	0	1	0	
Link4	0	1	1	1	
....					

<표 4> Link-Link Matrix

	Link1	Link2	Link3	Link4	....
Link1		2	0	1	
Link2	2		1	3	
Link3	0	1		1	
Link4	1	3	1		
....					

Step4 인 군집화 결과 대표 키워드는 <표 5>와 같다.

<표 5> 군집 대표키워드

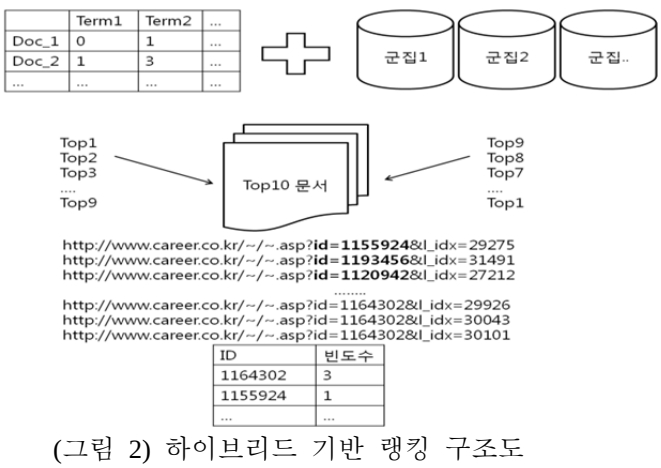
군집	대표키워드	군집	대표키워드
C1	신입/경력	C13	전문가(외환, 창구텔러, 베이커리개발)
C2	신재생에너지(IT 기술 융합)	C14	사무기술직
C3	경영혁신	C15	계약직, 연장공고, 특수직, 공무원
C4	모바일, 인증, 빌링, 게임	C16	마케팅, 브랜드
C5	영업	C17	재무/회계
C6	전략기획	C18	웹, 아이폰, 디지털음악서비스
C7	정기채용	C19	부문별
C8	유통, 신사업기획	C20	전문경력사원(전력, 전지, 발전부)
C9	IT	C21, 24	대기업 회사명 군집
C10	사업부(사업부별 모집군)	C22	대졸신입사원
C11	플랜트, Auto Cad, Battery	C23	2010년 하반기
C12	해외, 물류, 무역, 항공여행	C25	2011년 경력사원

Step5 인 키워드 매칭은 (그림 2)와 같다.

단계 1: 키워드 기준으로 단어 빈도수가 제일 높은 문서를 선별하고 이 문서들의 Doc_ID 을 수집하여 벡터에 삽입한다. 각 Doc_ID 의 빈도수를 구한 다음 높은 순서대로 문서를 추출한다.

단계 2: 키워드 기준으로 군집대표키워드와 일치될 경우, 해당군집에서 문서를 추출한다. 일치되지 않은 경우, TF-IDF 가중치가 제일 높은 군집 순으로 클릭빈도수가 제일 높은 문서를 추출한다.

단계 3: 군집 기법과 Doc_ID 을 기반으로 하여 추출된 문서를 동적인 비율에 따라 Top-10 문서를 웹브라우저를 통해 보여준다.



(그림 2) 하이브리드 기반 랭킹 구조도

4. 실험 및 결과

취업정보업체인 S 사가 회원들에게 발송한 취업메일링을 입력 자료로 사용한다. 2010 년 7 월부터 2011 년 6 월간 데이터를 수집하였고 이미지 URL 와 7 회 미만 클릭 사용자들은 제외시켰다. 추천리스트에 대한 성능을 평가하기 위해 Precision at Top-N 방법을 이용하였다[12]. 기존방법은 TF 기법이다.

4.1 실험 1 – 자바 개발자 입력

제안한 검색리스트의 Top1-3 은 “개발자” 키워드를 이용하여 군집에서 추출되었다. “자바” 키워드에 해당된 군집은 없었다. <표 6>에서와 같이 TF-IDF 가중치가 제일 높은 군집 순으로 추출하였고 군집대표키워드와 질의된 키워드를 통해 연관성이 입증된 바, 해당 군집에 대한 추출이 타당함을 확인할 수 있었다. 기존 검색 Top-1 인 “2011 년 UBIVELOX 대졸 신입/경력사원 공개 채용”의 경우 제안한 검색리스트에서는 Top-6 에 위치하고 있다. 기존 검색방식은 TF 기반인 반면 제안한 검색리스트의 경우 Doc_ID 빈도수에 의해 랭킹을 정한다.

<표 6> 자바 개발자 군집

추출된 군집ID	TFIDF	질의된 키워드	군집 대표키워드
C9	0.99	개발자	IT
C18	0.089	개발자	웹, 아이폰, 디지털음악서비스
C4	0.079	개발자	모바일, 인증, 빌링, 게임

4.1.1 기존 검색

2011년 UBIVELOX 대졸 신입/경력사원 공개채용
기술개발 부문 경력직 채용
2011년 SK C&C 경력 사원 채용
웹 개발자 / 다이어트 컨설턴트급 직원채용
개발직무 경력사원 공개채용
모바일사업본부 개발 경력/신입
기술연구소 신입 및 경력 모집
각부문 신입 및 경력 모집
웹프로그래머 경력사원 모집
2011년 경력직원 모집

4.1.2 제안한 검색리스트

S/W 개발자 채용
시스템&웹개발자 모집
웹 개발자 / 다이어트 컨설턴트급 직원채용
모바일사업본부 개발 경력/신입
2011년 SK C&C 경력 사원 채용
2011년 UBIVELOX 대졸 신입/경력사원 공개 채용
기술개발 부문 경력직 채용
개발직무 경력사원 공개채용
기술연구소 신입 및 경력 모집
각부문 신입 및 경력 모집

4.1.3 사용자그룹평가

현업에서 IT 개발직 업종에 종사하는 사용자들로 하여금 그룹으로 만들어 평가하였다. 각기 10 개의 리스트를 제공하고 사용자로 하여금 채용제목과 링크를 통해 해당 문서를 보고 각 Top-N 마다 평가하였다. Top-3 인 경우 기존이나 제안 모두 큰 차이는 없지만, Top-10 으로 갈수록 제안이 기존보다 나음을 <표 7>을 통해 알 수 있었다.

<표 7> 자바 개발자 기존/제안 평가

이름	Top-10	Top-7	Top-5	Top-3
김ㅇㅇ	4/5	3/5	3/3	3/3
유ㅇㅇ	7/8	4/6	3/5	2/2
조ㅇㅇ	6/7	4/6	3/5	2/2
김ㅇㅇ	5/5	3/5	2/3	2/2
김ㅇㅇ	6/7	3/5	2/3	2/2

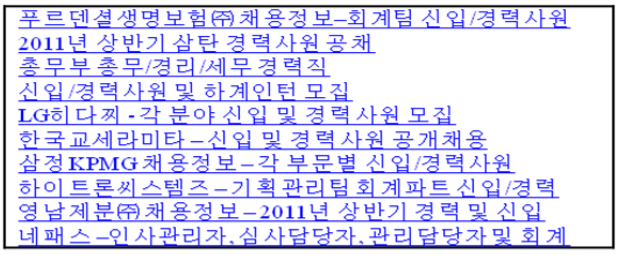
4.2 실험 2 – 재무 회계 입력

제안된 검색리스트의 Top1-6 의 경우 군집 C17 로부터 추출되었다. <표 8>와 같이 질의된 키워드와 군집 키워드와 일치되기 때문이다. 군집과 Doc_ID 빈도수기준으로 Top-10 문서를 6:4 비율로 추출하였다.

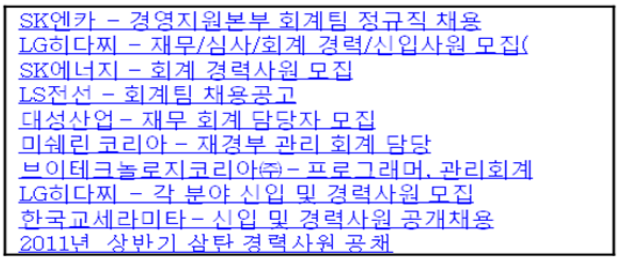
<표 8> 재무 회계 군집

추출된 군집ID	TFIDF	질의된 키워드	군집 대표키워드
C17	1.81	재무 회계	재무/회계
C4	0.35	재무 회계	모바일, 인증, 빌링, 게임
C23	0.25	재무 회계	2010년 하반기

4.3.1 기존 검색



4.3.2 제안한 검색리스트



4.3.3 사용자그룹평가

현업에서 재무 회계직 업종에 종사하는 사용자들로 하여금 그룹으로 만들어 평가하였다. 재무 회계직 업종의 경우 Top-3, 5, 7, 10 모두 기존 보다 제안이 높음을 <표 7>을 통해 알 수 있었다.

<표 7> 사무 회계 기존/제안 평가

이름	Top-10	Top-7	Top-5	Top-3
이ㅇㅇ	5/7	3/6	2/5	1/3
엄ㅇㅇ	4/7	3/5	3/5	2/3
윤ㅇㅇ	6/8	5/6	3/5	1/3
정ㅇㅇ	5/7	3/6	2/5	1/3
김ㅇㅇ	4/7	3/6	2/5	1/3

5. 결론 및 향후과제

실제 모든 카테고리에 적용해 보지는 못해서 자세한 통계치는 얻을 수 없었지만 군집을 이용한 검색기법이 기존보다 높음을 명확히 알 수 있었다. 또한 군집기반 사용시 군집 대표 키워드의 TF-IDF 값이 높을수록 추천된 문서뿐만 아니라 평가 또한 좋음을 알 수 있었다. 군집을 이용하여 그룹을 나눌 때의 메커니즘을 향상시켜 좀 더 매칭도가 높은 문서를 추출할 수 있도록 해야겠다. 군집이 아닌 다른 기법도 적용해 보고 실제 서비스화 할 수 있도록 구현 해야 겠다. 또한 검색 엔진에서 사용되는 질의어 유형을 분석하여 자연언어 질의어 처리기의 설계 및 구현도 필요하다고 생각한다.

참고문헌

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza. “Query recommendation using query logs in search engines”. In International Workshop on Clustering Information over the Web (ClustWeb, in conjunction with EDBT), Creete, Greece, March, Springer,LNCS, pp 588-596 , 2004.

[2] Z. Zhiyong, N. Olfa. “Mining Search Engine Query Logs for Query Recommendation”. In Proceeding WWW 2006

[3] S. Wedig and O. Madani. “A large-scale analysis of query logs for assessing personalization opportunities”. In Proceedings of KDD ’’06, pp 742-747, 2006.

[4] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. “Exploring Folksonomy for Personalized Search”. In Proc. of Search. In Proc. of SIGIR’ 08, pp 155-162, 2008.

[5] Dou, Z., Song, R., and Wen, J.R. “A large-scale evaluation and analysis of personalized search strategies”. In Proceedings of WWW ’ '07, pp 581-590, 2006.

[6] 용환승, 나연묵, 박종수, 승현우, 이민수, 이상준, 최린, “데이터마이닝”, 인피니티북스, pp 494-512, 2007

[7] 고모리 유스케, “프로가 되기 위한 웹기술 입문”, 위키북스 , pp14~83, 2012

[8] 배경만, 고영중, 최호섭, 김종훈. “논문 검색을 위한 K-means 알고리즘 기반의 검색 결과 내 Clustering”, 한국정보과학회 학술 심포지움 논문집 제 1 권 제 1 호, 2007

[9] 이정훈, 전서현. “검색 질의 확장을 위한 인기도 기반 단어 가중치 측정”, 정보과학회 소프트웨어 및 응용 제 37 권 제 8 호, 2010

[10] 김성민, 이성진, 이수원. “고속연관규칙을 이용한 문맥광고에서의 콘텐츠 추천”, 한국정보과학회 학술발표논문집 Vol.33 No.2B, 2006

[11] 한국어 분석 모듈, URL: http://nlp.kookmin.ac.kr/HAM/kor/index.html, 2012

[12] 이희춘. “협력적 필터링에서 개선 알고리즘을 이용한 Top-N 순위에 관한 연구”, The Korean data analysis society J Korean Data Anal Soc vol.9 no.1, 2007