도메인 온톨로지를 이용한 개인화된 개념기반 검색 기법

A Personalized Concept-based Retrieval Technique Using Domain Ontology

문현정(Mun Hyeon Jeong)*, 이수진(Lee Soo Jin)**,
김영지(Kim Young Ji)***, 우용태(Woo Yong Tae)****

초 록

본 논문에서는 도메인 온톨로지를 사용하여 개인화된 개념 기반의 검색 기법을 제안하였다. 제안 모델은 도메인 온톨로지를 이용한 컨텐츠의 대표 개념 추출, 컨텐츠 가중치와 개념 가중치를 이용한 사용자 프로파일 구성 그리고 개인화된 개념 기반 검색 과정으로 구성된다. 컨텐츠의 대표 개념은 *TScore* 기법을 이용하여 추출하였고, 사용자 프로파일은 개인 정보 수집 모듈을 통해 개념 가중치가 높은 개념을 대상으로 구성하였다. 개념 기반 검색을 위해 사용자 프로파일의 개념 집합과 컨텐츠의 대표 개념 집합간에 유사도를 비교하여 개인이 선호하는 개념의 우선순위에 의해 컨텐츠를 검색하였다. 본 논문에서 제안한 기법의 효율성을 검증하기 위하여 인터넷 사이트에서 컨텐츠를 수집하고 사용자 프로파일을 구성하여 실험하였다. 실험 결과, 제안한 검색 기법이 기존의 키워드 기반의 검색 기법에 비해 우수함을 보였다. 제안된 기법은 개인화된 추천 시스템이나 전자 도서관 등과 같은 분야에서 효율적으로 적용할 수 있으리라 기대된다.

ABSTRACT

We propose a personalized concept-based retrieval technique that uses domain ontology. Proposed system consist of representative concept extraction, user profile construction, and concept-based retrieval stages. First, we extract representative concept with using technique form contents and create the domain ontology. We compose user profile analysis that uses domain ontology for personalized concept-based retrieval. To verify the efficiency of the proposed technique, we perform experiment for Internet site in the engineering area. The results of experiment show that the proposed technique using the domain ontology and user profiles is more efficient than the existing techniques. Hence, the proposed concept-based retrieval technique can be expected to contribute to the development of an efficient personalized recommendation system or e-Commerce system.

키워드 : 온톨로지, 개념 기반 검색, 사용자 프로파일
Ontology, Concept-based Retrieval, User Profile

이 논문은 2004년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2004-050-D00016).

* 창원대학교 연구교수
** 창원대학교 일반대학원 컴퓨터공학과
*** 하이브레인넷 책임연구원
*** 창원대학교 컴퓨터공학과 교수

1. 서 론

기존의 검색 시스템에서 많이 사용되고 있는 키워드 기반의 검색 기법은 질의어와 컨텐츠로부터 추출된 키워드를 비교하여 검색하는 기법으로 여러 가지 문제점을 가지고 있다. 첫째, 사용자가 입력하는 키워드는 사용자마다 서로 다른 의미를 가질 수 있다. 둘째, 사용자에게 친숙하지 않은 도메인에서 적절한 키워드를 선정하기 어렵다. 셋째, 컨텐츠에 포함된 의미적인 개념은 키워드로 검색하기 어렵다[9, 13, 14].

이러한 키워드 기반 검색 기법의 문제점을 개선하기 위하여 메타데이터나 온톨로지를 이용한 검색 기법에 대한 연구가 진행되고 있다. 먼저, 메타데이터 기반의 검색 기법은 컨텐츠의 의미적인 검색을 위하여 컨텐츠에 메타데이터를 추가하여 검색하는 기법이다. 하지만 이 기법은 사용자의 질의와 메타데이터간의 의미적인 연결이나 추론이 어렵다[7]. 또한 대량의 컨텐츠에 대해 관리자가 수동적으로 메타데이터를 관리하는데 시간과 비용이 많이 소요된다[7, 12].

온톨로지를 메타데이타로 이용한 검색 기법은 온톨로지에 의해 구성되는 개념과 개념간의 관계를 이용하여 컨텐츠에서 대표 개념을 추출하여 사용자의 질의와 의미적으로 연관된 컨텐츠를 검색하는 기법이다[3, 6, 11, 18]. 국내에서도 온톨로지의 상·하위 개념 또는 동의어를 이용하여 질의어를 확장하여 검색하는 기법[1, 4]과 시소러스와 관련 개념을 이용하여 검색에 이용하는 연구가 진행되었다[2].

최근에는 사용자 프로파일을 이용한 개인화된 검색 기법에 대한 연구가 진행되고 있다[15-17]. 사용자 프로파일은 개인 정보와 함께 컨텐츠에 대한 개인적인 선호도 정보로 구성된다. 하지만 기존 연구에서는 사용자가 초기 사용자 프로파일을 명시적으로 구성하거나 키워드 기반으로 구성하는 관계로 컨텐츠에 대한 개인적인 선호도를 적절하게 반영하기 어려운 문제점이 있다.

본 논문에서는 도메인 온톨로지를 이용하여 개인화된 개념기반 검색 기법을 제시하였다. 본 연구는 크게 도메인 온톨로지를 이용한 컨텐츠별 대표 개념 추출, 컨텐츠 가중치와 개념 가중치를 이용한 사용자 프로파일 구성, 도메인 온톨로지 기반으로 사용자 프로파일을 확장한 개념기반 검색 과정으로 구성된다.

먼저, 도메인 온톨로지는 인터넷 사이트의 컨텐츠를 수집하여 생성하였다. 그리고 컨텐츠를 대표하는 개념 집합은 문현정 등이 제안한 *TScore* 기법을 이용하여 추출하였다[3]. 사용자 프로파일은 사용자가 명시적으로 입력한 개인정보와 사용자가 선호하는 상위 N개의 관심 개념으로 구성된다. 개인별 관심 개념은 본 연구에서 제안한 컨텐츠 가중치*(DW : Document Weight)*와 개념 가중치를 이용하여 구성하였다. 컨텐츠 가중치*(CW : Concept Weight)*는 컨텐츠별로 사용자의 액션 정보를 묵시적인 평가 기법에 의해 수집하였다. 개념 가중치는 사용자가 열람한 컨텐츠 중에서 관심 개념을 포함하는 컨텐츠 가중치에 의해 구성하였다.

개념 기반 검색은 사용자 프로파일의 개념 집합과 컨텐츠의 대표 개념 집합간에 유사도를 비교하여 개인이 선호하는 개념의

우선순위에 의해 컨텐츠를 검색하는 과정이다. 검색 과정에서 도메인 온톨로지를 이용하여 개인별 관심 개념에 대한 상·하위 개념과 연관 개념을 이용하여 개인별 관심 개념에 대한 검색 범위를 확장하였다.

본 논문에서 제안한 개념기반 검색 기법의 효율성을 검증하기 위하여 인터넷 사이트에서 수집한 컴퓨터, 정보통신, 산업공학 분야의 컨텐츠를 대상으로 실험하였다. 실험 결과, 본 논문에서 제안한 기법이 기존의 키워드 기반의 검색 기법보다 우수한 검색 효율을 보였다. 본 연구에서 제안한 검색 기법은 개인화된 추천시스템이나 전자상거래, 전자도서관 등과 같은 분야에 적용할 수 있으리라 기대된다.

2. 관련연구

2.1 키워드 기반의 검색 기법

키워드 기반의 검색 기법은 컨텐츠와 질의어간의 단순한 키워드 매칭에 의해 검색된 결과를 사용자에게 제공하는 기법이다. 하지만 이 기법은 단순히 컨텐츠에서 출현한 단어를 기준으로 검색하는 관계로 사용자가 키워드를 직접 선정하여 입력해야 하는 문제점이 있다[8, 9, 13, 14, 20]. 키워드 기반 검색 기법에서 사용자가 평균적으로 입력하는 단어의 수는 2.21개정도이다[5]. 또한 사용자들이 입력하는 키워드의 길이가 더욱 짧아지고 있으며 함축적인 의미를 포함하는 경향이 있다[5, 19]. 최근에 기존의 검색 기법의 문제점을 개선하기 위하여 메타데이터나 온톨로지를 이용한 검색 기법에 대한 연구가 진행되고 있다.

2.2 메타데이터 기반의 검색 기법

메타데이터 기반의 검색 기법은 컨텐츠를 대표하는 키워드를 컨텐츠에 추가하여 의미적인 검색을 하기 위한 기법이다. 메타데이터 기반의 검색 기법에 관한 연구는 David Vallet과 VICODI Project가 진행되었다[7, 12]. David Vallet 모델에서는 컨텐츠별로 메타데이터를 추가하여 검색하는 모델로 컨텐츠의 개념과 가중치가 부여된 키워드간의 유사도를 비교하여 검색 결과를 제공한다[7]. VICODI Project는 컨텐츠에 메타데이터를 추가하여 온톨로지 기반의 검색 시스템과 질의를 확장하기 위한 모델이다[12]. 하지만 이 기법은 메타데이타를 수동적으로 구성하는 관계로 메타데이타에 대한 관리 비용이 높은 문제점이 있다.

2.3 온톨로지 기반의 검색 기법

온톨로지 기반의 검색 기법은 온톨로지를 구성하는 개념과 개념간의 관계를 이용하여 검색하는 기법이다. 이 기법에서는 온톨로지의 계층 구조간의 관계를 정의하고 관계를 이용한 의미적인 검색을 지원한다. 온톨로지 기반의 검색 기법에 관한 연구는 TAP과 SEWISE 시스템 등이 진행되었다. TAP는 미국의 스탠포드 대학에서 SUO (Standard Upper Ontology) 온톨로지와

Cyc upper 온톨로지를 이용하여 컨텐츠에 대한 검색 영역을 확장하기 위해 제안된 시스템이다. 하지만 TAP에서는 RDF 기반의 RDQL 형태로 표현된 질의문을 처리해야 하는 문제점이 있다[18]. SEWISE는 웹상의 다양한 형태의 자원을 XML 형식으로 구조화하고 도메인 온톨로지를 기반으로 검색하기 위한 시스템이다. 하지만 이 시스템은 웹상의 다양한 자원들을 Wrapper를 이용하여 XML 형식으로 변환해야 하는 문제점이 있다[11].

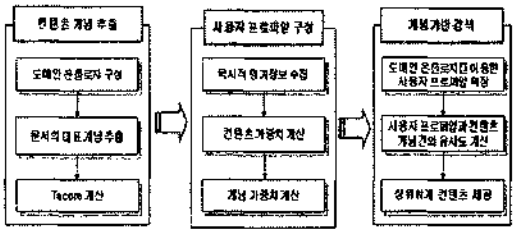
2.4 개인화된 검색 기법

개인화된 검색 기법은 컨텐츠에 대한 개인의 선호도를 반영하여 검색 범위를 줄이거나 출력 우선 순위를 조정하기 위한 기법이다. 개인의 선호도는 사용자가 명시적으로 입력하거나 사용자의 행동을 묵시적으로 분석하여 사용자 프로파일로 구성된다. 최근에 온톨로지 기반으로 사용자 프로파일을 구성하여 개인화된 문서 관리나 검색에 이용하기 위한 연구가 진행되고 있다. Castells 등은 사용자가 선호하는 문서에 포함된 용어로부터 연관된 키워드를 분석하여 사용자 프로파일을 구성하였다[17]. 사용자가 열람한 컨텐츠를 기준으로 온톨로지를 구성하고, 온톨로지에 열람한 컨텐츠를 매핑하여 관심 개념을 분석하기 위한 연구도 제시되었다[15]. 그리고 사용자 질의어를 분석하여 질의어에 사용된 개념과 개념의 빈도수를 분석하여 사용자 프로파일에 반영한 연구도 진행되었다[16].

3. 도메인 온톨로지를 이용한 개인화된 개념기반의 검색 기법

본 논문에서는 도메인 온톨로지를 이용하여 개인화된 개념기반 검색 기법을 제시하였다. 본 연구는 크게 도메인 온톨로지를 이용한 컨텐츠별 대표 개념 추출, 컨텐츠 가중치와 개념 가중치를 이용한 사용자 프로파일 구성, 도메인 온톨로지 기반으로 사용자 프로파일을 확장한 개념기반 검색 과정으로 구성된다.

<그림 1>은 본 논문에서 제안한 개인화된 개념 기반 검색 기법의 전체적인 개념도이다.



<그림 1> 개인화된 개념 기반 검색 기법 개념도

3.1 도메인 온톨로지를 이용한 컨텐츠의 대표 개념 추출

본 연구에서 컨텐츠를 대표하는 개념을 추출하기 위하여 문현정 등이 제안한 기법을 이용하였다. *TScore* 기법은 컨텐츠에서 대표 개념을 추출하기 위하여 도메인 온톨로지의 개념별 연관 용어 집합과 컨텐츠에서 추출한 대표 용어간의 유사도를 비교하기 위한 기법이다. 이 기법은 컨텐츠의 대표

용어와 도메인 온톨로지의 개념별 연관 용어 집합간의 유사도를 비교하여 임계치를 초과하는 *TScore* 값을 가지는 용어를 해당 컨텐츠에 대한 대표 개념으로 추출한다[3].

임의의 도메인 D_i를 대표하는 용어 dt_i에 대한 연관 용어 집합을 $AT_{dtj}=\{(at_{j1}, atw_{j1}), (at_{j2}, atw_{j2})\cdots,(at_{jl}, atw_{jl})\}$라 정의하면 대표 용어 dt_j에 대한 *TScore*는 식 (1)과 같이 정의된다.

$$TScore(dt_j)=\frac{1}{N}\sum_{k=1}^{0} atw_{jk}\times I_k \qquad (1)$$

여기서, $I_k=\sum_{l=1}^{P} ew_l\times atf^{i}_{lk}$

atw_{ik} : 임의의 대표 용어 dt_j에 대한 용어 at_k의 가중치

ew_l : 엘리먼트 e의 가중치

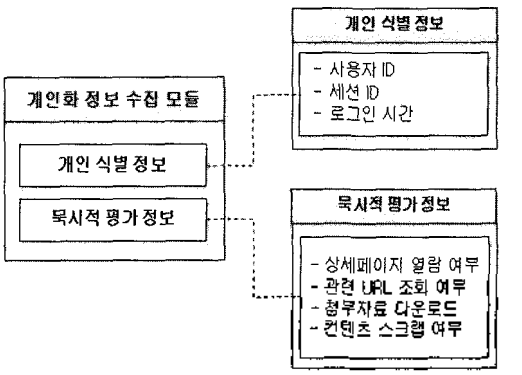
atf^{i}_{jm} : 엘리먼트 e_l에서 임의의 대표 용어 dt_j에 대한 용어 at_k의 빈도수

N : 임의의 대표 용어 dt_j에 대한 연관 용어 수

3.2 사용자 프로파일 구성

3.2.1 개인화 정보 수집

사용자 프로파일은 로그파일 수집 에이전트를 이용하여 로그파일로부터 개인 식별정보와 묵시적 평가 정보와 같은 개인화 정보를 동적으로 수집하여 구성된다. 〈그림 2〉는 개인화 정보를 수집하는 과정에 대한 개념도이다.



〈그림 2〉 개인화 정보 수집 과정

〈표 1〉은 사용자가 열람한 컨텐츠에 대한 개인화 정보를 수집한 예이다. 컨텐츠에 대한 개인별 선호도는 상세 정보에 대한 스크랩 여부와 관련 URL 조회 여부 등의 액션에 따라 식별하고 액션별로 가중치를 부여하여 구성하였다. 〈표 1〉에서 사용자 ID, 세션 ID 그리고 로그인 시간은 사용자 프로파일 구성을 위한 개인 식별정보이다.

〈표 1〉에서 홍길동이 열람한 컨텐츠 44916의 대표 개념은 '정보처리', 'MS SQL'이고, 컨텐츠 45492의 대표 개념은 'DBA', 'Oracle'이다. 그리고 홍길동이 열람한 컨텐츠에서 상세정보를 스크랩하거나 관련 URL을 조회

〈표 1〉 개인화 정보 수집 예

사용자 ID	세션 ID	로그인 시간	액션구분	컨텐츠ID	대표 개념
홍길동	DHGGMDILLMHB	2007/06/01	view	44916	정보처리, MS SQL
홍길동	DHGGMDILLMHB	2007/06/01	view	45492	DBA, Oracle
홍길동	DHGGMDILLMHB	2007/06/01	Scrap	45492	DBA, Oracle
홍길동	DHGGMDILLMHB	2007/06/01	url	45492	DBA, Oracle
…	…	…	…	…	…

하는 액션을 통해 해당 컨텐츠에 대한 관심도가 높다는 것을 알 수 있다.

3.2.2 컨텐츠 가중치 분석

컨텐츠 가중치 분석은 개인이 열람한 컨텐츠에 대한 액션의 종류와 빈도수에 따라 컨텐츠에 대한 개인별 선호도를 분석하는 과정이다. 컨텐츠에 대한 액션의 종류는 컨텐츠 열람, 스크랩, 다운로드, 관련 URL 조회 등으로 이루어진다. 컨텐츠에 대한 개인별 선호도는 컨텐츠에 대한 액션과 빈도수를 분석하여 컨텐츠 가중치로 추출하였다. 컨텐츠 집합 $D=\{d_1, d_2, \cdots, d_Q\}$에 대하여 컨텐츠에 대한 액션의 종류와 액션별 가중치를 $A=\{(a_1, aw_1), (a_2, aw_2), \cdots, (a_R, aw_R)\}$라 한다. 여기서 임의의 컨텐츠 d_k에 대해 n번의 액션이 발생한 경우, 컨텐츠 가중치 DW는 식 (2)와 같이 정의된다.

$$DW_{d_k} = \sum_{i=1}^{n} aw_i \qquad (2)$$

aw_i : 임의의 컨텐츠 d_k에 대해 i번째 액션 가중치

3.2.3 개념 가중치 분석

개념 가중치 분석은 컨텐츠에 포함된 개념에 대한 개인별 관심도를 분석하는 과정이다. 사용자가 열람한 컨텐츠에 포함된 개념 집합을 $UC=\{uc_1, uc_2, \cdots, uc_T\}$라 하면, 임의의 개념 C_h에 대한 개념 가중치 CW는 식 (3)과 같이 정의된다.

$$CW_{C_h} = \frac{\sum_{j=1}^{N} dw_{d_k}}{N} \qquad (3)$$

dw_{d_k} : 임의의 대표 C_h를 포함하는 컨텍스 d_k에 대한 컨텐츠 가중치

N : 열람한 컨텐츠 중 임의의 개념 C_h가 나타난 컨텐츠의 수

3.2.4 사용자 프로파일 구성

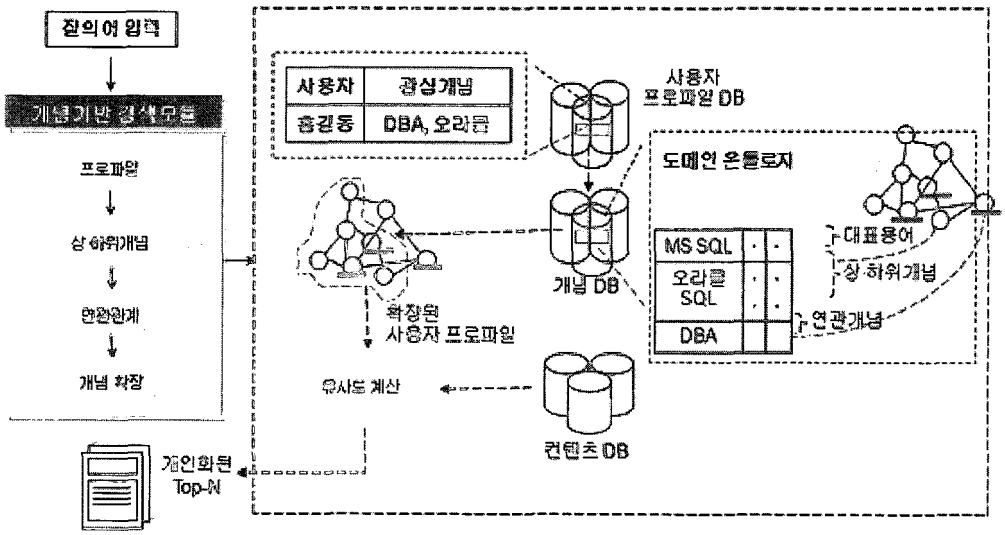
사용자 프로파일은 사용자에 대한 인적사항과 식 (3)에 의해 구성한 컨텐츠 가중치에 따라 추출한 개인별 상위 N개의 관심 개념으로 구성된다. 〈표 2〉는 사용자 프로파일을 구성한 예이다.

〈표 2〉 사용자 프로파일 구성 예

구 분	속 성	속 성 값
인적사항	이 름	홍길동
	전자우편	abc@changwon.ac.kr
	학 력	석사
	세부전공	컴퓨터/정보통신/산업공학
관심개념	관심개념	개념 가중치
	PHP	1.25
	RDBMS	1.25
	.NET	1.0
	C#	1.0

3.3 개인화된 개념기반 검색

본 연구에서 제안한 개인화된 개념기반 검색 기법은 단순 키워드 매치에 의해 검색결과를 출력하는 기존의 키워드 검색 기법과 다르게 개인별로 관심있는 컨텐츠를 개념적으로 검색하기 위한 기법이다. 〈그림 3〉은 본 연구에서 제안한 개인화된 개념 기반 검색 기법을 통하여 개인화된 상위 N개의

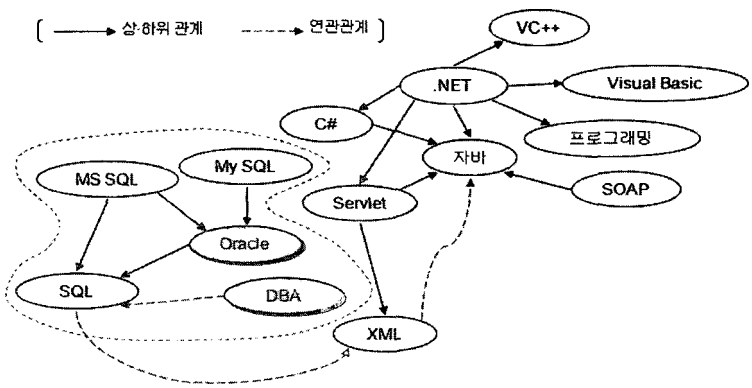


〈그림 3〉 개인화된 개념기반 검색기법

컨텐츠를 검색하여 출력하는 개념도이다.

본 연구에서는 개념 검색의 효율을 높이기 위하여 도메인 온톨로지를 이용하여 사용자 프로파일에 저장된 관심 개념을 중심으로 상·하위 관계와 연관 관계에 있는 개념을 포함하여 개념 검색 범위를 확장하였다. 〈그림 4〉는 'Oracle', 'DBA' 개념을 중심으로 1차적으로 연결된 도메인 온톨로지를 이용하여 개념을 확장한 예이다.

개인화된 개념 기반 검색은 사용자 프로파일에 저장된 개인별 관심 개념과 컨텐츠에서 추출한 대표 개념과 유사도 비교를 통하여 검색 결과를 출력하는 과정이다. 개념 기반 검색을 위하여 유사도를 비교하는 과정은 다음과 같다. 먼저, 사용자 프로파일을 구성하는 개념 집합을 $C=\{(c_1, cw_1), (c_2, cw_2),$



〈그림 4〉 도메인 온톨로지를 이용한 개념 검색 범위 확장 예

…, (c_m, cw_m)}라 하고, 임의의 개념 c_i에 대해 상·하위 또는 연관 관계에 있는 개념으로 확장한 집합과 컨텐츠에서 추출한 개념 집합을 $EC=\{(ec_{i1}, ecw_{i1}), (ec_{i2}, ecw_{i2}),\cdots,(ec_{iN}, ecw_{iN})\}$ 그리고 컨텐츠에서 추출한 개념 집합을 $D=\{d_1, d_2\cdots,d_0\}$라 한다. 식 (4)는 도메인 온톨로지를 이용하여 개인별 관심 개념을 확장한 집합과 컨텐츠에서 추출한 개념 간에 유사도를 계산하는 식이다.

$$similarity(EC, D)=\sum_{i=1}^{M}(\sum_{j=1}^{N} f(ec_{ij}, dc_k)\times cw\times ecw_{ij} \quad (4)$$

$f(ec_{ij}, dc_k)$: 확장한 사용자 프로파일의 개념 ec_{ij}와 컨텐츠 개념 dc_k를 비교하는 함수

여기서, $f(ec_{ij}, dc_k)$는 확장한 개념이 컨텐츠내에 존재하면 1을 반환하고, 존재하지 않으면 0을 반환하는 함수이다.

4. 실험 및 결과

본 논문에서 제안한 개념기반 검색 기법의 효율성을 검증하기 위하여 인터넷 사이트의 사용자를 대상으로 로그파일을 수집하여 실험하였다. 실험에서는 일주일 동안 794명의 사용자가 493건의 컨텐츠에 대하여 열람한 16,365건의 웹 로그를 분석하여 사용자 프로파일을 구성하였다.

4.1 도메인 온톨로지 생성

도메인 온톨로지는 컴퓨터/정보통신/산업공학 관련분야의 컨텐츠 3,765건을 대상으로 생성하였다. 전체 컨텐츠에서 추출된 기본 용어에 대하여 전처리 과정을 거친 후 최종적으로 추출된 전문용어는 585개이다. 도메인 온톨로지는 대표용어와 관계를 정의하고 대표용어별 연관용어 집합을 구성하여 생성

〈표 3〉 대표용어별 연관용어 집합

대표용어	연관용어 집합
.NET	Visual Basic, C++, 프로그래밍, Visual C++, 자바, C#
CBD	설계, Consultant, 분석, 아키텍처, Engineering, Solution ERP, CMM
CORBA	Network, 임베디드, 자바, PROTOCOL, Visual C++, LINUX
DBA	Network, 자바, Program, Oracle, 구축, 서버관리
EJB	JSP, LINUX, XML, 설계, Solution, EDI
ERP	프로그래밍, Solution, CBD, 자바, MIS, Delphi, Enterprise
KMS	지식경영, 산업공학, 분석, Package, SI, Data Mining, Enterprise
Oracle	My SQL, Applet, MS SQL, 자바, DBA, JSP, SQL, MFC, DirectX
XML	SOAP, Servlet, Applet, EDI, 자바, EJB, SQL, MS SQL, UML
유비쿼터스	WIBRO, 멀티캐스트, 미들웨어, Wireless
정보보호	저작권, 정보보안, DRM, Firewall, MIS, PKI

〈표 4〉 개인화 정보 수집 결과

사용자ID	액션 유형					
	채용정보ID	상세정보 열람	관련URL 조회	채용정보 스크랩	첨부파일 다운로드	DW
홍길동	45427	2	0	0	0	2
	45442	6	8	0	0	14

하였다. 〈표 3〉은 컴퓨터/정보통신 분야의 대표용어별 연관용어 집합을 구성한 예이다.

4.2 사용자 프로파일 구성

사용자 프로파일을 구성하기 위해 개인화 정보 수집 모듈을 통해 개인별 개념 선호도를 묵시적인 평가정보 수집 기법에 의해 구성하였다. 개인별 개념 선호도는 식 (2)의 컨텐츠 가중치와 식 (3)의 개념 가중치를 구한 후, 개념 가중치가 높은 상위 N개의 개념으로 구성하였다. 〈표 4〉는 개인이 열람한 컨텐츠에 대하여 액션 유형에 따라 식 (2)에 의해 컨텐츠 가중치를 구성한 예이다.

〈표 5〉은 사용자 프로파일에서 식 (3)에 의해 개념 가중치를 계산하여 상위 5개의 개념을 개인별 관심 개념으로 구성한 예이다.

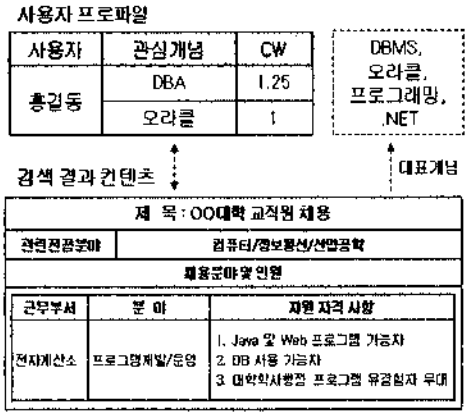
4.3 개인화된 개념기반 검색 실험

개인화된 개념기반 검색은 사용자 프로파일과 온톨로지의 개념간의 관계를 이용하여 온톨로지의 개념을 확장하고 컨텐츠와 유사도 계산을 통해 상위 N개의 검색 결과를 제

〈표 5〉 개인별 관심 개념 구성 예

사용자ID	관심 개념	개념 가중치
홍길동	PHP	1.25
	RDBMS	1.25
	.NET	1
	C#	1
	DBMS	1
성춘향	.NET	1.67
	PHP	1.67
	ActiveX	1.67
	ERP	1.67
	웹프로그래밍	1.67

공하는 과정이다. 홍길동의 사용자 프로파일에 저장된 관심 개념은 'DBA', '오라클'이며, 컨텐츠에서 도메인 온톨로지와 TScore 기법에 의해 추출된 대표 개념은 'DBMS', '오라클', '프로그래밍', '.NET'이다. 〈그림 5〉는 홍길동의 관심 개념을 이용하여 식 (4)의 유사도 계산에 의해 개인화된 개념 검색 결과를 출력하는 과정을 보여주는 그림이다.



〈그림 5〉 도메인 온톨로지를 이용한 개인화된 개념기반 검색 결과

4.4 제안기법의 효율성 비교

본 연구에서 제안한 기법의 성능을 평가하기 위하여 평균절대오차(Mean Absolute Error, MAE)와 recall, precision, F-measure 평가 기법에 의해 검색 효율성을 비교하였다.

MAE는 통계적인 정확도를 측정하는 방법의 하나로 컨텐츠에 대한 사용자의 평가 정보와 수치화된 예측 값을 비교하기 위해 사용된다. 식 (5)는 MAE에 대한 정의식이다.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \quad (5)$$

recall은 사용자가 선호하는 컨텐츠 중에서 얼마나 많은 컨텐츠가 검색되었는지 나타내는 평가 방법이다. 식 (6)은 recall에 대한 정의식이다. 높은 recall 값은 사용자가 선호하는 컨텐츠가 검색되었다는 것을 의미한다. 하지만 사용자가 선호하지 않은 컨텐츠도 검색될 가능성이 높다.

$$recall = \frac{\text{정확하게 검색된 컨텐츠 수}}{\text{사용자가 선호하는 컨텐츠 수}} \quad (6)$$

precision은 검색 결과 중에서 몇 개의 컨텐츠를 사용자가 실제로 선호하는지를 나타내는 평가방법이다. 식 (7)은 precision에 대한 정의식이다. 높은 precision 값은 검색된 모든 컨텐츠가 사용자에게 정확하게 검색되었다는 것을 의미한다. 하지만 사용자가 선호하는 컨텐츠가 검색되지 않을 가능성도 있다.

$$precision = \frac{\text{정확하게 검색된 컨텐츠 수}}{\text{검색 리스트의 컨텐츠 수}} \quad (7)$$

이러한 recall과 precision 값은 상호 보완적인 관계가 있으므로 적절한 조정 과정이 필요하다. Lewis 등은 recall과 precision의 문제점을 보완하기 위하여 recall과 precision을 결합한 F-measure 개념을 제안하였다. 식 (8)은 F-measure에 대한 정의식이다.

$$F_\beta = \frac{(\beta^2+1) \times precision \times recall}{\beta^2 \times precision \times recall} \quad (8)$$

〈표 6〉 제안기법과 기존시스템의 효율성 비교

구 분	MAE	recall	precision	F-measure
기존시스템	0.728	0.963	0.184	0.553
제안기법	0.329	0.454	0.246	0.737

〈표 6〉은 본 연구에서 제안한 기법과 키워드 기반 검색 기법간에 MAE, recall, precision, F-measure 평가기법에 의해 검색 효율을 비교한 결과이다. 실험 결과에서처럼 제안 기법이 모든 평가 항목에서 우수한 결과를 보였다. 키워드 기반의 검색 결과는 사용자의 선호도와 상관없이 키워드와 일치하는 컨텐츠를 많이 포함하는 관계로 MAE, recall 값은 높게 나타나고, precision, F-measure 값은 낮게 나타났다.

5. 결 론

본 논문에서는 도메인 온톨로지를 이용하여 개인화된 개념기반 검색 기법을 제시하였다. 본 연구에서 제안한 개인화된 개념기반 검색 기법은 단순 키워드 매칭에 의해 검색 결과를 출력하는 키워드 기반의 검색 기법과 다르게 개인별로 관심있는 컨텐츠를 개념적으로 검색할 수 있는 새로운 기법이다. 제안 기법은 키워드를 중심으로 프로파일을 구성하거나 특정 관계만을 확장하는 기존의 검색 기법과는 다르게, 온톨로지의 모든 관계를 이용하여 개념을 확장하여 검색 효율을 높일 수 있다. 또한 본 기법은 묵시적인 평가정보 수집 기법에 의해 개인별 행동 패턴에 따라 동적으로 사용자 프로파일을 구성할 수 있다.

먼저, 인터넷 사이트에서 컨텐츠를 수집하여 컨텐츠를 대표하는 대표용어 집합을 추출하고, 용어간의 상·하위 개념과 연관 개념을 이용하여 도메인 온톨로지를 생성하였다. 그리고 본 연구에서 제안한 컨텐츠 가중치와 개념 가중치를 이용하여 개인별 관심 개념을 추출하여 사용자 프로파일을 구성하였다. 컨텐츠 가중치는 컨텐츠별로 사용자의 액션 정보를 묵시적인 평가 기법에 의해 수집하였다. 개념 가중치는 사용자가 열람한 컨텐츠 중에서 관심 개념을 포함하는 컨텐츠에 의해 구성하였다.

개념 기반 검색은 사용자 프로파일에 저장된 개념 집합과 컨텐츠에서 추출한 개념 집합간에 유사도에 의해 개인이 선호하는 개념을 포함하는 컨텐츠를 출력하는 과정이다. 본 연구에서는 검색의 효율성을 높이기 위하여 도메인 온톨로지를 이용하여 개인별 관심 개념에 대한 상·하위 개념과 연관 개념을 이용하여 개인별 관심 개념에 대한 검색 범위를 확장하였다.

본 논문에서 제안한 검색 기법의 효율성을 검증하기 위하여 인터넷 사이트의 사용자를 대상으로 실험 데이터를 수집하여 MAE, recall, precision, F-measure 평가기법에 의해 검색 효율을 비교하였다. 실험결과, 본 연구에서 제안한 기법의 성능을 평가하기

위하여 기존 검색 기법보다 우수한 성능을 나타내었다. 앞으로 본 연구에서 제안한 검색 기법은 개인화된 추천시스템이나 전자상거래, 전자도서관 등과 같은 분야에 적용할 수 있으리라 기대된다.

참 고 문 헌

[1] 강민구, 온톨로지를 이용한 시맨틱 질의어 확장과 검색, 숭실대학교 석사학위논문, 2005.

[2] 고상일, 개념기반 검색시스템에서의 개념 자동생성, 중앙대학교 석사학위논문, 2003.

[3] 문현정, 우용태, "지식 문서에서 도메인 온톨로지를 이용한 개념 추출 기법," 정보처리학회 논문지, 제13권, 제3호, pp. 309-316, 2006.

[4] 임수연, 송무희, 이상조, "온톨로지내의 계층관계를 이용한 문서검색," 춘계학술발표회 논문집, 제31권, 제1호, pp. 646-649, 2004.

[5] B. Jansen, A. Spink andT. Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," Information processing & Management, Vol. 36, pp. 207-227, 2000.

[6] B. M. Fonseca, P. Golgher, and B. Possas, "Concept-Based Interactive Query Expansion," CIKM Workshop, pp. 696-703, 2005.

[7] D. Vallet, M. Fernandez and P. Castells, "An Ontology-Based Information Retrieval Model," LNCS 3532, pp. 455-470, 2005.

[8] F. Liu, C. Yu, W. Meng, S. Binghamton and A. Chowdhury, "Effective keyword search in relational databases," Proc. Int'l Conf. on ACM SIGMOD, pp. 563-574, 2006.

[9] F. Tanudjaja and L. Mui, "Persona: A Contextualized and Personalized Web Search," Proc. Int'l Conf. on System Sciences, Vol. 3, pp. 53-61, 2002.

[10] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti and S. Sudarshan, "Keyword Searching and Browsing in Databases using BANKS," Proc. Int'l Conf. on Data Engineering, pp. 431-440, 2002.

[11] G. Gardarin, H. Kou, K. Zetourni, X. Meng and H. Wang, "SEWISE: An Ontology-based Web Information Search Engine," Proc. Int'l Conf. on NLDB, pp. 106-119, 2003.

[12] G. Nagypal, "Improving information retrieval effectiveness by using domain knowledge stored in ontologies," COOPIS, pp. 780-789, 2005.

[13] J. A. Royo, E. Mena, J. Bernad and A. Illarramendi, "Searching the Web: From Keywords to Semantic Queries," Int'l Conf. on Information Technology and Applications, pp. 244-249, 2005.

[14] J. Fiaidhi, S. Mohammed, J. Jam and

A. Hasnah, "A Standard Framework for Personalization Via Ontology-Based Query Expansion," Parkistan Journal of Information and Technology2, pp. 96-103, 2003.

[15] M. Grcar, D. Mladenic and M. Grobelnik, "User Profiling for Interest-focused Browsing History," Workshop on User Aspects of the Semantic Web, pp. 99-109, 2005.

[16] P. K. Bhowmick, S. Sarkar, S. Sarkar and A.Basu, "Acquisition of User Profile for Domain Specific Personalized Access," Int'l Conf. on Information Technology, pp. 20-23, 2004.

[17] P. Castells, M. Fernandez, D. Vallet, P. Mylonas and Y. Avrithis, "Self-Tuning Personalized Information Retrieval in an Ontology-Based Framework," LNCS 3762, pp. 977-986, 2005.

[18] R. Guha, R. Mccool and E. Miller, "Semantic Search," Proc. Int'l Conf. on WWW, pp. 700-709, 2003.

[19] S. Y. Rieh and H. Xie, "Patterns and Sequences of Multiple Query Reformulations in Web Searching: A Preliminary Study," Proc. on 64th ASIST Annual Meeting, pp. 246-255, 2001.

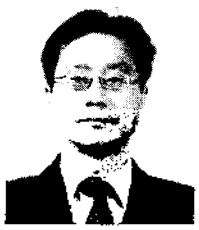
[20] V. Hristidis andY. Papakonstantinou, "Discover: Keyword Search in Relational Databases," Proc. Int'l Conf. on 28th VLDB, pp. 670-681, 2002.

저 자 소 개

문현정 (E-mail : mun@changwon.ac.kr)
1994. 한국방송대학교 전자계산학과 졸업 (이학사)
1996. 창원대학교 일반대학원 전자계산학과 졸업 (이학석사)
2003. 창원대학교 일반대학원 컴퓨터공학과 졸업 (공학박사)
2004 ~ 현재 창원대학교 연구교수
관심분야 KDD, 온톨로지마이닝, 시맨틱웹

이수진 (E-mail : gomi97@nate.com)
2002. 진주산업대학교 컴퓨터공학과 졸업 (공학사)
2007. 창원대학교 일반대학원 컴퓨터공학과 졸업 (공학석사)
관심분야 KDD, 온톨로지마이닝

김영지 (E-mail : yjkim@hibrain.net)
1997. 창원대학교 전자계산학과 졸업 (이학사)
1999. 창원대학교 일반대학원 전자계산학과 졸업 (이학석사)
2004. 창원대학교 일반대학원 컴퓨터공학과 졸업 (공학박사)
2004 ~ 2007. 고신대학교 초빙교수
2007 ~ 현재 하이브레인넷 책임연구원
관심분야 온톨로지마이닝, 추천모델, e-Learning



우용태 (E-mail : ytwoo@changwon.ac.kr)
1982. 경북대학교 전자공학과 졸업 (공학사)
1984. 경북대학교 일반대학원 전자공학과 졸업 (공학석사)
1995. 경북대학교 일반대학원 전자공학과 졸업 (공학박사)
1987 ~ 현재 창원대학교 컴퓨터공학과 교수
관심분야 데이터마이닝, 온톨로지마이닝, 시맨틱웹