

# 웹 정보 검색 이력을 이용한 사용자 의도 자동 추출

박기남<sup>†</sup> · 정순영<sup>††</sup> · 서태원<sup>††</sup> · 지혜성<sup>†</sup> · 이태민<sup>†</sup> · 임희석<sup>††</sup>

## 요 약

본 논문은 사용자가 정보 욕구를 정확하게 질의어로 입력하고, 원하는 정보가 검색될 수 있도록 지원하기 위한 사용자 의도 자동추출과 이를 이용한 인텐션 맵 구축 방법을 제안한다. 제안하는 방법은 동일한 검색어를 입력한 사용자들의 검색 이력 데이터를 이용하여 사용자 의도 자질을 선정하고, 클러스터링 알고리즘과 사용자 의도 추출 알고리즘을 이용하여 사용자 의도를 추출하였다. 추출된 사용자 의도는 지식표상 이론에 근거한 인텐션 맵으로 표현하였다. 제안한 인텐션 맵의 효용성 분석을 위하여 현재 국내 상용 검색엔진에서 제공받은 2,600개의 사용자 검색 이력 데이터를 이용하였다. 실험결과 인텐션 맵을 이용한 정보검색이 일반 검색엔진을 이용 할 때 보다 통계적으로 유의미한 만족도를 나타내었다.

주제어 : 웹 정보검색 시스템, 사용자 의도, 인텐션 맵

## Automatic Extract User Intention from Web Search Log

Kinam Park<sup>†</sup> · Soonyoung Jung<sup>††</sup> · Taewon Suh<sup>††</sup> · Hyesung Ji<sup>†</sup> ·  
Taemin Lee<sup>†</sup> · Heuiseok Lim<sup>††</sup>

## ABSTRACT

This paper proposes a method to extract a user's intention automatically and implementation of intention map that support a user can appropriate search results using a user' information need accurately. It selects user intention based on searching history obtained from previous users' same queries and extracts user intentions by using clustering algorithm and user intention extraction algorithm. Extracted user intentions are represented in an intention map base on a theory of knowledge representation. For the efficiency analysis of intention map, we extracted user intentions using 2,600 search history data which provided by a current domestic commercial search engine. The experimental results using the information intention map search when using general search engines represent more than satisfaction was statistically significant.

Keywords : Information Retrieval System, User Intention, Intention Map

<sup>†</sup> 정 회 원: 고려대학교 컴퓨터교육과

<sup>††</sup> 종신회원: 고려대학교 컴퓨터교육과

논문접수: 2008년 10월 30일, 심사완료: 2008년 11월 23일

\* 교신저자: 임희석 (limheiseok@korea.ac.kr)

\* 본 논문은 교육과학기술부 한국연구재단의 뇌과학 기술 연구 프로그램(2009-0093899)의 지원으로 수행되었음

## 1. 서 론

인터넷 정보의 폭발적인 증가와 인터넷 호스트의 증가에 따라 정보의 양은 기하급수적으로 늘어나고 있다. 하지만 웹 문서의 정보량에 따라 인덱스 하는 웹 정보 검색 시스템의 인덱스 페이지 비율은 오히려 감소하고 있다. 때문에 정보의 신뢰도는 낮아지고, 웹 정보 검색 시스템을 이용하여 검색자가 원하는 정보를 찾기 위한 노력과 비용이 증가되고 있다. 그래서 검색자가 원하는 웹 문서의 검색은 매우 중요한 기술이다.

기존 웹 정보 검색 시스템들은 검색 결과를 도출할 때, 검색자로부터 입력받은 질의어에 포함된 키워드와 연관된 문헌 집합을 빈도정보를 이용하여 추출하였다[1]. 이러한 방법은 검색자의 검색 의도를 반영한 질의어의 의미가 아니라 질의어에 포함된 단어의 형태에 따른 빈도 가중치를 이용한 방법이기 때문에 사용자의 검색의도를 정확하게 파악할 수 없는 문제점을 갖고 있다.

검색자의 만족도를 높이기 위하여 최근 웹 정보 검색 시스템 연구는 사용자의 검색 의도를 파악하거나 검색결과를 효과적으로 전달하기 위한 연구 영역으로 확대되고 있다[6]. 즉, 초기의 웹 정보 검색 시스템은 정보 공유를 중요하게 생각하여 검색자가 원하는 정보를 단순 키워드 매칭이나 개략적인 연관성에 따라 웹 문서에서 대량으로 추출하는 것만을 고려하였다면, 최근에는 방대한 정보가 내재되어 있는 문서에서 사용자가 원하는 정보를 정확히 파악하여, 추출하는 것을 중요하다고 보는 것이다.

일반적으로 검색에 사용되는 질의어는 길이가 짧고, 의미가 중의적이며 함축적이다[13][14]. 때문에 검색자의 의도가 구체적으로 표현되지 않은 짧은 질의어를 사용한 검색 결과는 나쁠 수밖에 없다. 정보검색 학술회의인 TREC에서도 초기에 100여 단어로 이루어진 질의어를 사용하다가 비현실적이라는 지적 때문에 평균 15개 단어로 제안하였다[4]. 특히 질의어가 동형이의어나 이형 동의어일 경우, 어떤 의미로 사용됐는지, 동의어 혹은 유사어가 어떤 것인지 알아야 하기 때문에 양질의 검색결과를 얻기 힘들다. 이과 같은 문제

점을 해결하기 위한 대표적인 방법으로 질의확장(query expansion)과 연관 검색어(relevant keyword)가 있다[5].

질의확장은 적합성 피드백(relevance feedback)과 시스템에 기반을 둔 적합성 피드백에 의한 방법으로 나뉜다. 적합성 피드백 방법은 검색자가 직접 검색 결과를 바탕으로 연관 문서와 비연관 문서를 판단하여, 질의어를 다시 작성하고 검색함으로써 좀 더 좋은 검색 결과를 낼 수 있는 방법이다. 하지만 검색자의 지식수준, 해당 질의어에 대한 이해정도, 웹 정보 검색 시스템의 알고리즘에 대한 이해여부에 따라 확장되는 질의어가 달라질 수 있고, 정확도가 보장되지 않는다. 이를 보안한 방법이 시스템에 기반을 둔 적합성 피드백을 이용한 방법이다. 이 방법은 초기 검색결과 상위 N개를 검색자의 의도와 연관된 문서라고 판단하고, 자동으로 적합성 피드백을 이용한 검색결과에 포함시켜 검색자가 확장한 질의어 검색결과에 반영하는 방법이다. 하지만 이 방법 역시 초기 검색 결과가 좋지 못할 경우 성능이 오히려 크게 떨어지는 단점이 있다.

연관 검색어는 검색자의 질의어를 분석하여 기존 검색자들의 질의어 이력에서 유사한 검색 질의어들을 선정하여 검색자에게 추천하는 방법이다. 하지만 이 방법은 주어진 질의어를 구성하고 있는 단어와 가장 밀접한 단어를 선택할 때 사용빈도에만 의존할 수밖에 없는 단점이 있다.

웹 정보 검색 시스템을 이용해 검색자에게 정확한 정보를 제공하기 위해서는 사용자의 요구를 정확하게 파악하는 문제 해결이 선행 되어야 한다. 그러면 사용자의 요구를 어떻게 파악할 수 있을까? 이 물음에 대한 답을 위해 기존 연구에서는 웹 정보 검색 시스템에 입력되는 질의어에 포함된 핵심 단어를 사용자 요구로 가정 하였지만, 본 연구에서는 기존 검색자들의 검색 결과 이력을 바탕으로 한 사용자 요구 분석 방법을 제안한다. 본 연구에서는 정확한 사용자 요구 분석을 위해 검색자 요구를 다음과 같이 정의한다.

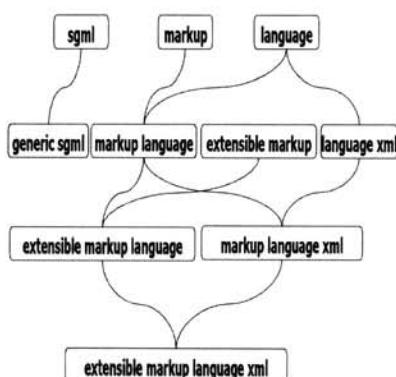
검색자 요구 : 정보 검색자의 정보 수요를 정확하게 만족시킬 수 있는 개인적, 상황적 의미를 말하는 포괄적 의미.
---

- 검색하고 싶은 것
- 의미적 연관성이 있는 단어집합
- 정보수요자의 탐색 대상을 가장 잘 표현하는 단어집합
- 정보수요자의 탐색 욕구를 명확하고 세세하게 표현하는 개인적/상황적 의미

본 연구는 앞에서 정의한 검색자 요구를 바탕으로 검색 성능 향상을 위해 사용자가 원하는 의도 정보를 자동으로 추출하고, 사용자에게 제공할 수 있는 방법을 제안한다.

## 2. 관련 연구

웹 정보 검색 시스템을 이용하는 개인 검색자들의 질의어는 암묵적으로 서로 다른 검색 의도를 포함할 수 있기 때문에 중의성이 크다. 또, 일반적으로 질의어가 짧기 때문에 충분한 사용자 의도 파악을 위한 정보량을 갖지 못하는 문제점이 있다. 이를 해결하기 위해 검색 알고리즘과 정보의 신뢰를 강화하려는 전통적인 검색 시스템 연구방법 이외에도 사용자의 검색 의도를 파악하기 위한 연구 영역으로 확대되고 있다.



<그림 1> 계층적 질의확장

[2]는 <그림 1>과 같이 구 (phrase) 기반의 벡터모델을 통해 사용자 의도를 계층적으로 확장하고, 이를 기준으로 SVM(support vector machine)을 이용해 긍정적(positive) 혹은 부정적(negative) 문서로 분류하고, 사용자에게 의도에 적합한 문서

를 추천하는 방법을 제안하였다. 하지만, 모델은 사용자 피드백을 통해 질의어를 확장하고, 패턴을 고려하였기 때문에 사용자 참여가 이루어지지 않을 경우 성능에 영향을 줄 수 있고, 초기 질의어 패턴 학습에 대한 내용이 고려되지 않았다.

[3]는 사용자 질의어의 핵심 단어 특징과 개념 특징을 이용한 NBC(naïve bayes classifier) 모델을 구현하였다. 모델은 사용자의 다음 검색 행위를 예측 하도록 학습되었으며, 사용자의 의도를 도출하는 질의어 핵심 단어 및 개념과 같은 언어적 특성을 바탕으로 이루어졌다. 같은 의미의 다른 용어일지라도 다른 의도를 반영하기 때문에 이 연구에서는 핵심 단어로부터 의도되는 위계적 개념을 사용하였으며, 워드넷(WordNet)으로부터 어휘 개념을 여러 형태로 변형하여 적용시켰다. 실험을 통해 질의어를 구성하는 핵심 단어의 개념 특징을 사용하는 것이 효과적인 검색결과를 도출할 수 있는 것으로 나타났으며, 84%의 정확도를 보였다.

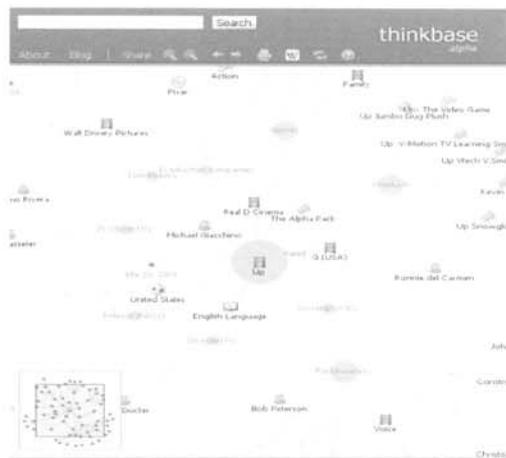
현재 사용자 검색의도 파악을 통해 서비스되고 있는 시스템들 중에서 원더휠(Wonder Wheel, <http://google.com>)은 질의어와 관련된 유사 질의어를 <그림 2>와 같이 방사형으로 제시한다.



<그림 2> Wonder Wheel

사용자들은 검색하고자 하는 관련 질의어를 입체적으로 확인할 수 있다. 방사형 그래프에서 제공되는 유사 질의어는 최대 8개 제시되고 있고, 모든 질의어에 대해 자동으로 제시되고 있으나, 각 유사 질의어들 간의 유사도에 대한 정의가 명확하지 않고, 빈도정보에 의존하고 있다.

유사한 형태로 University of Auckland의 Thinkbase 서비스가 있다(<http://thinkbase.cs.auckland.ac.nz>). 공개형 온톨로지 저작 데이터베이스인 freebase에 기반을 둔 시맨틱 네트워크를 제공한다.



<그림 3> Thinkbase

질의어를 입력하면 그와 연관된 정보들이 <그림 3>과 같이 방사형 그래프 형태로 표시되며, 정보들 간의 연관관계가 표시된다. 정보는 wikipedia(<http://www.wikipedia.org>)와 같이 집단 지성으로 구축된 온톨로지에 기반을 두어 제공된다. 하지만 질의어에 대한 추상적인 내용을 배제하고, 객관적인 사실만을 제공하기 때문에, 다양한 사용자의 검색 의도 충족에는 한계를 갖는다.

이와 같이 사용자 의도 파악을 위한 많은 연구가 이루어지고 있고, 현재 서비스 되고 있지만, 정보 검색자의 정보 욕구를 정확하게 만족시키지 못하고 있다. 대부분의 연구들이 정보검색자의 개인적/상황적 의미를 고려하지 않고, 질의어의 형태학적 연관성이나 빈도정보에 의존하고 있기 때문이다.

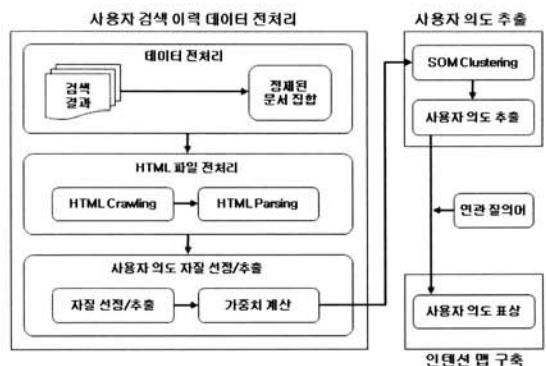
### 3. 인텐션 맵(Intention Map) 구축

인텐션 맵이란 서론에서 정의한 검색자의 검색 요구에 부합할 수 있는 정보의 집합으로 질의어와 관련된 연관키워드 집합과 사용자 의도를 나타내는 핵심단어 집합으로 구성된다. 본 논문은 정보 검색자가 검색 의도를 효과적으로 입력하고,

정보 수요에 적합한 검색 결과를 얻을 수 있도록 지원하기 위한 사용자 의도 자동 추출과 이를 이용한 인텐션 맵 구축 방법을 제안한다. <그림 4>는 본 논문이 제안한 인텐션 맵 구축 방법을 구현한 시스템 구성도이다.

제안하는 방법은 사용자 의도를 자동으로 추출하기 위하여 검색자가 입력한 질의어와 동일한 질의어를 이용하여 과거에 검색 작업을 수행한 사용자들의 검색 이력 데이터를 활용한다. 특정 질의어에 대한 검색 결과를 분류하고, 각 부류가 사용자의 의도를 포함하고 있는 데이터 집합이라 간주한다. 사용자 의도는 각 부류를 분석하여 부류내의 중심 단어들로 설정되며, 설정된 사용자의 의도는 연관 검색어와 함께 인텐션 맵 구축에 사용된다.

**인텐션 맵:** 동일한 의미의 질의어를 사용하는 정보 검색자들의 정보수요를 정확하게 만족시킬 수 있는 상황적 의미 집합.



<그림 4> 시스템 구성도

#### 3.1 사용자 검색 이력 데이터 전처리

웹 정보 검색 시스템 사용자들의 검색 이력은 검색자의 검색 행위를 사실적으로 반영한다[8]. 검색 이력 중에서도 검색 결과는 사용자 각자의 검색 욕구를 충족시키기 위한 결과물이기 때문에 검색자의 직/간접적인 의도를 반영한다.

### 3.1.1 사용자 검색이력 데이터

현 정보검색 시스템들은 사용자 검색 결과를 URL형태로 저장한다[6]. 또한 검색결과로 출현한 웹 문서는 한글 외에 영문자, 한자, 숫자, 특수 문자, 문장 부호, 그림 등 다양한 형태의 콘텐츠를 포함하고 있다. 본 연구에서는 신뢰성 있는 사용자 검색 의도 추출을 위해 다음 전처리 과정을 통해 제안하는 방법의 복잡성을 감소 시켰다.

1. 일정 빈도 이하의 값을 가지는 URL을 모두 제거 한다(사용자들이 선택한 중복된 검색 결과는 사용자 의도를 다수 포함할 수 있지 만, 상대적으로 낮은 빈도의 검색 결과는 사용자 의도 정보량이 작다. 따라서 일정 빈도 이하의 URL을 제거한다).
2. URL에 링크된 에러 페이지를 삭제한다(잘못 된 URL이나, 링크된 문서를 삭제한다).
3. 검색결과를 보여주는 검색 시스템의 검색결과를 제거한다(검색결과를 보여주는 문서는 사용자 질의어에 대한 정보를 포함하지 않고, 질의어에 대한 URL결과만을 나타내기 때문에 사용자 의도 정보가 포함 되었다고 볼 수 없다).

사용자 검색결과 URL 데이터를 전처리 과정을 통해 정제 하고, URL에 해당되는 실제 웹 문서를 수집한다. 수집된 문서는 사용자 의도 정보를 포함한 단어뿐만 아니라 HTML Tag, Java Script, Link Page 주소 등 다른 많은 정보로 구성되기 때문에 이를 제거한 후 질의어와 관련된 문자만을 추출한다.

### 3.1.2 사용자 의도 자질 선정 및 가중치 계산

사용자 의도 자질 선정 및 가중치 계산 과정에서는 정제된 검색이력 데이터(웹 문서)에서 사용자 의도 추출을 위한 자질을 선정하고, 각 자질의 상대적 중요도를 빈도정보를 이용하여 계산한다. 일반적으로 자질은 구문적 분리도가 높은 단어들로 선정하는데, 개념을 표현하는 명사와 고유명사

를 주로 사용 한다[10]. 본 연구에서는 명사를 자질로 선정하여 추출한다. 자질 선정 과정은 사용자 검색결과 문서에서 분석배제 과정을 통해서 명사로 분석될 가능성이 없는 문자(기호, 특수 문자 등)를 제거하고, 단일 형태소 및 후접어 분석을 통하여 명사를 추출한다. 두 과정 중 처리되지 않는 단어를 고려해서 명사를 자질로 추출한다.

사용자 의도 자질들의 상대적 중요도는 기존 정보검색 방법에서 많이 사용되고 있으며, 정보이론에 따라 정보량이 많은 단어에 가중치를 부여 할 수 있고, 계산량이 적은 장점이 있는 TF\*IDF 값을 사용한다[9].

$$w_{ij} = tf_{ij} \cdot \log\left(\frac{N}{df_i}\right) \quad <\text{수식 } 1> \text{ 가중치 계산}$$

가중치  $w_{ij}$ 는 사용자 의도를 포함하고 있는  $i$ 번 째 검색 결과 데이터의  $j$ 번째 자질의 가중치를 말한다.  $tf_{ij}$ 는  $i$ 번째 데이터에서  $j$ 번째 자질이 나타난 횟수를 말하며,  $N$ 은 검색 결과 전체 데이터 개수를,  $df_j$ 는  $i$ 번째 자질이 나타난 데이터의 수를 말한다. 즉,  $TF$ 는 데이터에서 빈도수가 높은 자질에 가중치를 많이 주며,  $IDF$ 는 반대로 여러 검색 결과 데이터에서 나타나는 자질의 가중치를 감소시킨다.

## 3.2 사용자 의도 분류

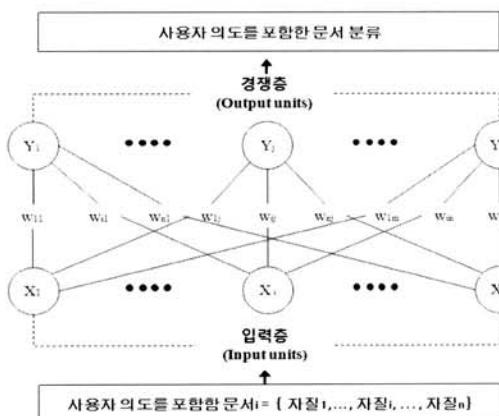
사용자 검색이력 데이터를 바탕으로 추출된 사용자 의도 자질을 기준으로 검색이력 데이터를 분류함으로써 각 부류들은 사용자의 의도를 표현한 인텐션들 집합을 구성한다. 분류를 위해 본 논문에서는 클러스터링(clustering)을 기법을 사용한다. 클러스터링을 위한 알고리즘은 구현 방식에 따라 크게 계층적(hierarchical) 방법, 분해적(partitioning) 방법 그리고 응집화(clumping) 방법으로 구분된다[11]. 본 논문에서는 각 구현 방식을 대표 하는 알고리즘들(SOM, K-means, EM, FarthestFirst, Cobweb) 중에서 사용자 의도 자질 분류 위한 도메인 의존적(domain dependent)인 SOM(Self Organizing Map) 클러스터링 알고리즘을 사용 한다.

SOM 알고리즘은 다른 알고리즘들에 비해 웹 문서 분류에서 좋은 성능을 나타낸다. 본 연구에서는 실험을 통해 이를 검증하였다. 검증을 위해 우리는 웹에서 총 450건의 블로그(Blog)와 뉴스를 수작업으로 수집 및 분류(애완동물, 연애, 예절, 음식, 이민, 임신, 자동차, 주택, 패션)하였고, 각 알고리즘에 대한 성능을 측정하였다. 결과는 아래와 같다.

<표 1> 클러스터링 알고리즘 성능 평가

Algorithm	SOM	K-means	EM
Correct clustered instances	214 (47.6)	202 (44.9%)	85 (18.9%)
Incorrectly clustered instances	236 (52.2%)	248 (55.1%)	365 (81.1%)
	<b>Cobweb</b>	<b>FarthestFirst</b>	
	50 (11.1%)	55 (12.2%)	
	400 (88.9%)	395 (67.8%)	

SOM은 학습 단계에서 유사한 패턴끼리 2차원의 특징 지도를 조직화하여 영역 지도를 형성 한 후 인식 단계에서 이미 학습 단계에서 훈련된 연결 가중치하에서 미지의 특징 벡터에 대해서 경쟁 층에서 반응이 일어나는 위치를 통하여 부류를 만드는 알고리즘이다. 결과적으로 본 연구에서 추구하는 사용자의 의도에 대한 분류를 가능하도록 하는 알고리즘이다. SOM을 이용한 사용자 의도 분류 과정은 <그림 5>과 같이 표현할 수 있다.



<그림 5> 사용자를 의도를 포함한 문서 분류 과정

사용자 의도 자질을 포함한 검색 결과 데이터는 자질의 가중치 계산 과정을 통해 가중치 벡터로 표현되고, SOM 알고리즘의 입력으로 주어진다. SOM 알고리즘을 통해 사용자 의도를 포함한 데이터들은 유사한 사용자 의도 자질을 포함한 부류로 분류된다. SOM 알고리즘을 이용한 검색 이력 데이터 분류 과정을 상세히 기술하면 아래와 같다.

- 연결강도를 초기화함( $(W_{11}, \dots, W_{nm})$ )
- 새로운 입력 벡터를 제시함[사용자의도 포함 데이터]
- 입력 벡터와 모든 경쟁층 유닛들 간의 거리를 계산함.  
입력  $i$ 와 출력  $j$  사이의 거리  $D(j)$ 는 <수식 2>과 같이 계산함
$$D(j) = \sum_i (w_{ij} - x_i)^2 \quad <\text{수식 } 2>$$
- 여기서  $X_i$ 는  $i$ 번째 입력 벡터이고,  $W_{ij}$ 는  $i$ 번째 입력 벡터 와  $j$ 번째 출력 사이의 연결강도 임
- 최소 거리에 있는 출력을 선택함  
최소 거리  $D(j)$ 인 출력  $j_*$ 를 선택함
- 출력  $j_*$ 와 그 이웃들의 연결강도  $W_{ij}$ 를 재조정함
$$W_{ij}(\text{new}) = w_{ij}(\text{old}) + \alpha [x_i - w_{ij}(\text{old})] \quad <\text{수식 } 3>$$
- 단계 2로 가서 반복함

### 3.3 사용자 의도 추출

본 논문에서는 분류된 사용자 의도 포함 검색 결과 데이터들에서 부류를 대표하는 핵심 단어를 추출한다. 추출된 단어는 검색 사용자의 검색 의도를 반영한다. 핵심 단어 추출을 위한 방법은 부류 내의 데이터를 구성하고 있는 자질들의 가중치를 고려해서 부류의 대표 단어로 추출하는 방법과 부류들 간의 차이를 가장 잘 나타내는 대표 단어를 추출하는 방법을 생각할 수 있다. 하지만 본 연구에서는 성능 향상을 위해 두 가지 방법 모두를 고려한 형태의 알고리즘을 구현하였고, 이를 이용하여 사용자 의도를 추출한다. 사용자의도 추출 알고리즘은 아래와 같다.

**Algorithm Extract Intention (C)**

```

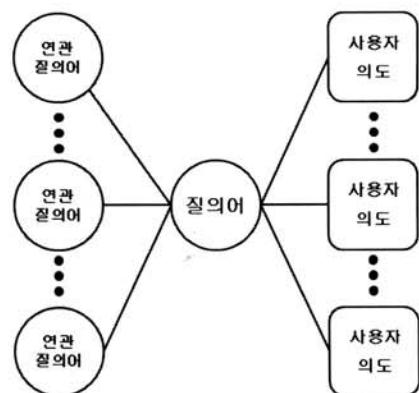
1   foreach( $c_i$  in  $C$ )
2     foreach( $t_j$  in  $c_i$ )
3        $new\_w(t_j) = \sum_{d \in c_i} \sum_j old\_w(t_j)$ 
4        $idf(t_j) = |\{d : t_j \in c_i\}|$ 
5        $ICF(t_j) = \frac{1}{|\{c : t_j \in C\}|}$ 
6        $Score(t_j) = new_w(t_j) \times idf(t_j) \times ICF(t_j)$ 
7   return high-scored terms

```

$C$ 는 부류들의 집합,  $c_i$ 는 부류,  $t_j$ 는 사용자 의도 자질,  $d$ 는  $i$ 번째 부류에 속한 데이터를 나타낸다.  $new\_w$ 는 부류  $c_i$ 에 속한 데이터 전체에 출현한 자질들의 가중치( $old\_w$ ) 합이다.  $idf$ 는 부류  $c_i$  내 자질 발생 데이터 수를,  $ICF$ 는 자질이 발생한 부류 개수의 역수이다.  $Score$ 는 부류 내 의미 있는 자질의 가중치를 나타내는 지표로서,  $new\_w$ 와  $idf$ ,  $ICF$ 의 곱으로 구할 수 있다. 최종적으로 각각의 부류에 대하여  $Score$  값이 높은 자질을 부류의 대표 자질로 선택한다.

### 3.4 인텐션 맵 표상

추출된 사용자 의도는 검색자가 입력한 질의어와 함께 맵(map) 형태로 표상(representation)된다. 맵은 사용자 의도를 적절하게 표현할 수 있도록 인지 심리학적 지식 표현 방법론인 스키마(schema) 모형을 사용하였다[7]. 스키마는 기억에 저장된 개념을 데이터 구조의 한 단위로 보고, 개념과 개념의 속성 간의 상호관계를 망 형태로 구성하고 있는 것을 말한다. 본 논문은 스키마를 사용자 질의어로 상정하고, 스키마의 속성을 사용자의 의도들의 집합으로 표현한다. <그림 6>은 인텐션 맵의 개념적 표상을 나타낸다.



&lt;그림 6&gt; 인텐션 맵

<그림 6>에서 연관 검색어는 대표 인텐션과 검색 질의어와의 연관관계 확장을 위해 아래와 같은 규칙을 사용하여 인텐션 맵에 추가 하였다.

1. Collaborative Filtering: 검색결과 문서와 질의어 간의 관계를 분석하여 검색결과 문서와 연관성이 높은 질의어를 추출한다.
2. Query Sequence: 질의어 입력 전/후에 입력 빈도가 높은 질의어를 추출한다.
3. Query Clustering: 질의어와 유사한 검색어 중 입력 빈도가 높은 질의어를 추출한다.

## 4. 실험 방법 및 결과

본 연구에서 제안한 검색자의 검색사용 이력을 이용한 사용자 의도 자동 추출 및 인텐션 맵 구축을 위한 실험 방법 및 결과는 아래와 같다.

### 4.1 실험데이터(사용자 검색이력 데이터)

제안하는 방법의 실험을 위해서 국내 상용 검색 시스템의 사용자 검색 이력 데이터를 사용하였다. 데이터는 사용자가 질의어에 따라 실제 열람한 웹 페이지를 의미한다. 데이터 전처리를 통해 실험에 사용한 데이터는 아래의 표와 같다.

&lt;표 2&gt; 검색어 별 클릭 URL 수

검색 질의어	URL 수(Web Page)	
	전처리 전	전처리 후
가방	552	527
꿈해몽	790	765
낚시용품	732	654
별자리	526	496
합계	2,600	2,442

총 2,600개의 사용자 검색 이력 데이터는 데이터 전처리 과정을 통해 1이하의 빈도를 갖는 URL, 오류 페이지 그리고 검색 시스템의 검색 결과가 제거되었고, 총 2,442개로 정제 되었다. 각 데이터에서 추출된 사용자 의도 자질의 개수는 아래 표와 같다.

&lt;표 3&gt; 검색어 별 클릭 URL 수

검색 질의어	사용자 의도 자질 수
가방	6,525
꿈해몽	5,857
낚시용품	4,094
별자리	2,547
합계	4,755

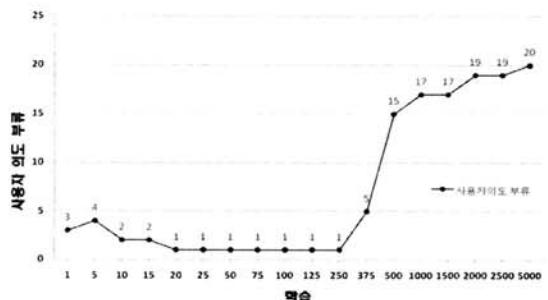
## 4.2 사용자 의도 분류

사용자 의도 분류를 위해 TF\*IDF 가중치 계산 방법을 사용하여 각 검색 질의어에 대한 자질들의 가중치를 계산하였다. 계산된 평균 4,755개의 사용자 의도 자질을 이용하여 SOM 알고리즘의 입력을 설계하였다. 아래 표는 입력설계 방식을 이용한 실제 예이다.

&lt;표 4&gt; 입력벡터 설계

사용자 의도 포함 데이터	Input vector			
	1	2	...	6,525
	가방	검색	...	판다면
1	0.58989	0.68937	...	0
2	0.76966	0.64203	...	0
3	0.53269	0.50000	...	0
4	0.66479	0.52838	...	0

실험 결과 평균 19개의 사용자 의도 부류 정보를 얻을 수 있었다. 부류 정보는 <그림 7>에서와 같이 학습데이터(training data)를 2000회 학습하였을 때부터 사용자 의도 부류 개수가 안정되는 것을 볼 수 있다. 분류된 부류는 검색 질의어에 대한 사용자 의도 인텐션들의 집합을 의미한다.



&lt;그림 7&gt; 학습 횟수에 따른 부류 개수의 변화량

## 4.3 인텐션 추출

인텐션 집합을 구성하는 의도들 가운데는 사용자 의도를 적절히 반영하지 못한 인텐션들이 있을 수 있고, 의도들 간의 사용자의 의도 정보의 차이가 있기 때문에 사용자 의도 추출 알고리즘을 이용해 대표 인텐션을 추출하였다. <표 5>는 질의어 별 추출된 대표 인텐션들의 일부이다.

&lt;표 5&gt; 검색어 별 대표인텐션

검색 질의어	대표 인텐션
가방	가방쇼핑몰, 가짜가방, 명품가방,...
꿈해몽	운세, 돼지꿈, 태몽,...
낚시용품	낚시터, 입질, 낚시인,...
별자리	별자리운세, 초신성, 사랑개론,...

## 5. 평가

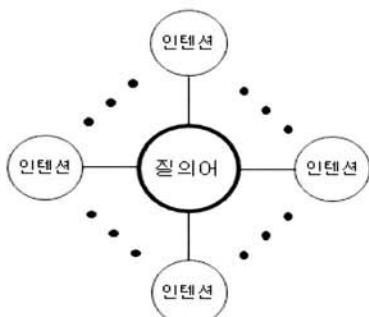
본 논문에서 제안한 인텐션 맵의 효용성 평가를 위해 행동실험을 실시하였고, 분석을 통해 사용자 검색의도 자동 추출 방법을 검증하였다. 행동실험은 인텐션 맵 표현도 검증과 만족도 검증으로 이루어졌다. 인텐션 맵 표현도 검증이란 사용자 의도 추출 과정을 통해 구축된 인텐션 맵이

사용자 검색의도를 적절하게 반영하고 있는 것에 대한 검증이다. 인텐션 맵 만족도 검증이란 검색의도를 내포한 과제 수행에 있어, 인텐션 맵을 이용한 검색에 대한 만족여부를 검증한 것이다.

## 5.1 인텐션 맵 표현도 검증

### 5.1.1 실험방법

실험은 대학생 57명(남자 37명, 여자 20명)을 대상으로 실시하였다. 실험을 위해 <그림 8>과 같이 4개의 주제어에 대해 각각 자동으로 추출된 사용자 검색의도를 스키마 모형의 표현 방식으로 인텐션 맵을 구축하였다. 실험에서는 검색어 질의어와 검색의도들 간의 관계성에 대해 검색의도들이 해당 ‘질의어를 얼마나 잘 표현하였는지’에 대한 인텐션 맵 표현도를 평가하였다.



<그림 8> 스키마모형에 적용된 인텐션들

표현도는 Likert scale로 측정하였다. Likert scale이란 1932년 R.리커트가 고안한 태도측정법으로 같은 종류의 내용에 관계되는 여러 의견을 모아 이를 3~7단계의 연속체 척도상의 점수에 맞추어 그 합계 점을 가지고 태도의 점수로 삼는 상가평정척도와, 항목분석에 의하여 작성하는 내적 일관성 척도를 결합시킨 태도측정법이다. 표현도는 ‘매우 부적절함’을 1점, ‘매우 적절함’을 7점으로 하여 1점 단위로 체크할 수 있도록 구성하였다.

### 5.1.2 실험결과

<표 6>과 같이 인텐션 맵 표현도에 대한 리커드척도 평균값은 5.39(최고: 꿈해몽 5.62, 최저: 가방 5.04)으로 각 질의어에 대한 검색의도를 비교적 적절하게 표현하였다고 평가되었다.

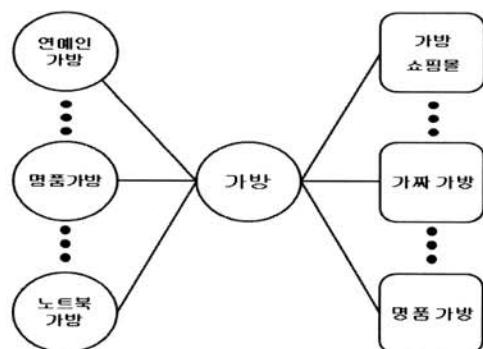
<표 6> 키워드 별 인텐션 맵 표현도

키워드	표현도
별자리	5.66
꿈해몽	5.62
가방	5.04
낚시용품	5.24
평균	5.39

## 5.2 인텐션 맵 사용자 만족도 검증

### 5.2.1 실험방법

실험은 대학생 37명(남자 24명, 여자 13명)을 대상으로 실시되었다. 이중 반응 데이터가 결손되거나 극단적인 반응을 한 7명의 자료는 4개 질의어 전체에 대한 분석과정에서 제외하였다. 실험을 위해 <그림8>과 같이 기 구축된 인텐션 맵에 연관 검색어를 추가한 메타 검색엔진을 구축하였다. 인텐션 맵 중앙에 질의어를 기준으로 좌측에는 연관 검색어를, 우측에는 인텐션들을 배치하였다.



<그림 8> 실험에 사용한 인텐션 맵의 예

실험참가자는 다음의 순서에 따라 검색과제 수행 및 만족도 평가를 실시하였다. 참가자는 선정된 4개의 질의어에 대해 인텐션 맵을 이용한 검

색과 일반검색을 실시하였다. 4개의 질의어에 대해 개인의 검색의도가 다를 수 있기 때문에 각 질의어에 대한 검색을 시작하기 전에 검색의도가 담긴 짧은 이야기식 지문을 제시하여 그 의도에 부합하는 웹페이지를 검색하도록 과제를 부여하였다.

1. 배경설문을 작성한다.
2. 검색의도를 생성하는 짧은 지문을 읽는다.
3. 인텐션 맵을 이용한 검색 또는 일반검색에 의해 산출된 검색결과 웹페이지들을 살펴보고 검색의도에 부합하는 웹페이지 순서대로 URL을 기재한다.
4. 검색결과에 대한 만족도와 일치도를 7점 척도로 평가한다. (만족도: 검색결과에 대해 어느정도 만족하십니까, 일치도: 원했던 검색결과와 어느정도 일치합니까)

### 5.2.2 실험결과

인텐션 맵을 이용한 검색과 일반검색의 만족도 점수에 대하여 paired t-test를 실시하였다. <표 7>와 같이 검증 결과 4개의 질의어에 대한 평균 만족도는 인텐션 맵을 이용한 검색이 5.56("만족", SD=.63)으로 일반검색이 4.43("보통", SD=.86)으로 실험참가자들은 일반검색에 비해 인텐션 맵을 이용한 검색결과에 대해 더욱 만족하였으며 이 차이는  $p<.05$ 에서 통계적으로 유의하였다 ( $t(29)=7.020$ ,  $p=.000$ ).

<표 7> 만족도 비교 평가 결과

		IM검색 평균(SD)	일반검색 평균(SD)	차이검증결과
가방	만족도	5.76 (1.19)	4.14 (1.60)	$t(36)=5.025$ , $p=.000$
낚시 용품	만족도	5.23 (1.38)	4.20 (1.86)	$t(29)=2.589$ , $p=.015$
꿈해몽	만족도	5.86 (1.08)	4.95 (1.41)	$t(36)=4.105$ , $p=.000$
별자리	만족도	5.45 (1.04)	4.60 (1.61)	$t(29)=2.589$ , $p=.015$
평균	만족도	5.56	4.43	

4개의 질의어 각각에 대한 분석에서도 인텐션 맵을 이용한 검색결과에 대한 만족도가 일반검

색에 비해 높았으며 이 차이는  $p<.05$ 에서 통계적으로 유의하였다.

## 6. 결론 및 향후 연구 방향

본 논문은 웹 정보검색 시스템의 사용자 만족도 향상을 위해 사용자 의도를 자동으로 추출하고, 사용자에게 제공할 수 있는 방법을 제안하였다. 사용자 의도 자질 추출을 위하여 과거 동일한 검색어를 이용하여 검색 작업을 수행했던 검색 사용자들의 질의어에 대한 검색 결과를 바탕으로 사용자 의도 자질을 선정하고, 클러스터링 알고리즘과 사용자 의도추출 알고리즘을 이용하여 사용자 의도를 추출하였다. 추출된 사용자 의도는 지식표상 이론을 바탕으로 인텐션 맵으로 표현하였다.

제안한 방법의 효용성 검증 실험은 현재 국내 상용 검색엔진에서 제공받은 2,600개의 사용자 검색 이력 데이터를 이용하였다. 이를 이용하여 사용자 의도를 추출하였고, 인텐션 맵을 구축하였다. 그리고 인텐션 맵을 이용한 메타 검색엔진을 구축하고, 행동 실험을 통해 본 연구에서 제안한 방법을 검증하였다. 검증결과 인텐션 맵을 이용한 검색에서 통계적으로 유의미한 결과를 나타내었다.

본 논문에서 제안한 사용자의 의도 추출과 인텐션 맵 구축 방법의 의의는 다음과 같다. 첫째, 사용자 검색의도 자동 분석이 가능하다. 사용자 질의어에서 나타나는 키워드 추출이 아닌 의미적 질의어(e.g., 형용사로 분류되는 경우: 재미있는, 최신, 가격이 싼 등의 질의, 직관적이지 못한 경우: 비만 종류, 다른 키워드를 입력해야 되는 경우: 빅뱅실험이 위험한 이유, 빅뱅실험 위험성, 찾고자 하는 내용을 전혀 모르는 경우: 컴퓨터 구매정보, 중의적인 의미를 갖는 경우 - 사과)를 사용자 검색이력에 기반을 두어 자동으로 추출할 수 있다. 둘째, 사용자의 검색 의도 자동 추출을 통하여 질의어 가이드를 제공할 수 있는 기반 연구가 가능하다. 기존의 검색 서비스에서는 사용자의 질의의 의도파악이 아닌 검색기의 성능을 높여 질의어와 문서간의 keyword가 가장 유사한 문서집합을 결과 값으로 제시하였으나, 사용자가 정확한 질의를 입력하지 못함으로써 검색 의도 파악이 불가능하였다. 셋째, 현재 운영 중인 검색엔진

의 실제 사용자 검색이력을 바탕으로 본 연구에서 제안한 방법을 실험함으로써 신뢰성 있는 실험 결과를 도출하였다.

향후에는 사용자 서비스 측면의 인텐션 맵 표상 또한 구조에 대한 추가적인 연구가 필요할 것이다. 본 연구에서 사용한 지식표상 구조는 이론적 모델에 기초하기 있기 때문에 실제 서비스 측면에서 사용자가 쉽게 사용할 수 있는 형태의 인텐션 맵 표상 구조가 개발되어야 할 것이다.

### 참 고 문 헌

- [1] Dreilinger, D. & Howe, A. E.(1997). Experience with selecting search engine using metasearch. *ACM Transactions on Information Systems*, 15(3), 195-222
- [2] GunWoo Park, JinGi Chae, Dae Hee Lee & SangHoon Lee.(2008). User Intention based Personalized Search : HPS(Hierarchical Phrase Serch). *the WSEAS International Conference on Applied Computing Conference*, 7, 205-210
- [3] Zheng Chen, Fan Lin, Huan Liu, Wei-Ying Ma & Liu Wenyin.(2002). User Intention Modeling in Web Applications Using Data Mining, *WorldWideWeb*, 5(2), 181-192
- [4] Voorhees, E.M., & Harman, D.(1997). Overview of the seventh Text Retrieval Conference (TREC-7). In *Proceedings of the 7th TREC Conference*, 1-23
- [5] Xu, J. & Croft, W.B.(1996). Query Expansion Using Local and Global Document Analysis. *Proc. ACM SIGIR Int'l Conf. Research and Development in Information Retrieval*, 4-11
- [6] Daniel E. Rose & Danny Levinson.(2004). Understanding user goals in web search. *Proceedings of the 13th international conference on World Wide Web*, May 17-20
- [7] Rumelhart, K.E.(1980). *Schemata: The building blocks of cognition*, N.J.: Erlbaum.
- [8] Jansen, B. & Spink, A.(2006) How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing & Management In Formal Methods for Information Retrieval*, 1(42), 248-263.
- [9] Salton, G. & McGill, M. J.(1983) *Introduction to Modern Information Retrieval*, McGraw-Hill
- [10] Mel'cuk, A.I.(1988). *Dependency Syntax: Theory and Practice*. State Univ. of New York Press
- [11] Gose, E., Johnsonbaugh, R. & Steve J. (1996). *Pattern Recognition and Image Analysis*. Prentice Hall
- [12] Fausett, V.(1994). *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall
- [13] Jansen, B.J., Spink, A. & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management*, 86(2), 207-227.
- [14] Spink, A.(2001). Searching the Web: The Public and Their Queries. *J. Am. Soc. Information Science and Technology*, 53(2), 226-234.



## 박 기 남

- 2004 천안대학교  
컴퓨터학과(학사)  
2006 한신대학교  
컴퓨터정보학과(석사)  
2006~현재 고려대학교  
컴퓨터교육학과 박사과정

관심분야: 자연어처리, 인지과학, 컴퓨터 교육  
E-Mail: spknn@korea.ac.kr

## 정 순 영



- 1990 고려대학교  
전산과학과(학사)  
1992 고려대학교  
전산과학과(석사)  
1997 고려대학교  
전산과학과 (박사)

2000~현재 고려대학교 컴퓨터교육과 교수  
관심분야: 데이터베이스, 컴퓨터교육  
E-Mail: jsy@comedu.korea.ac.kr



## 지 혜 성

- 2009 한신대학교  
컴퓨터공학과(학사)  
2008~현재 고려대학교  
컴퓨터교육학과  
석사과정

관심분야: 자연어처리, 컴퓨터 교육  
E-Mail: hyesung84@korea.ac.kr

## 임 희 석



- 1992 고려대학교  
컴퓨터학과(학사)  
1994 고려대학교  
컴퓨터학과(석사)  
1997 고려대학교  
컴퓨터학과 (박사)

2008~현재 고려대학교 컴퓨터교육과 교수  
관심분야: 컴퓨터교육, 자연어처리, 인지신경과학  
E-Mail: limhseok@korea.ac.kr



## 이 태 민

- 2008 고려대학교  
컴퓨터교육(학사)  
2008~현재 고려대학교  
컴퓨터교육학과

석사과정  
관심분야: 데이터마이닝, 컴퓨터 교육  
E-Mail: persuade@korea.ac.kr



## 서태원

- 1992 고려대학교  
전기공학과(학사)  
1995 서울대학교  
전자공학과(석사)  
1997 Georgia institute of  
technology (박사)

2008~현재 고려대학교 교수  
관심분야: 컴퓨터교육  
E-Mail: suhtw@korea.ac.kr