Asia Pacific Journal of Information Systems
Vol. 21, No. 1, March 2011

A New Approach to Automatic Keyword Generation Using Inverse Vector Space Model* **

Wonchin Cho***, Sangkyu Rho****, Jiyoung Agnès Yun*****, Jinsoo Park******

Recently, numerous documents have been made available electronically. Internet search engines and digital libraries commonly return query results containing hundreds or even thousands of documents. In this situation, it is virtually impossible for users to examine complete documents to determine whether they might be useful for them. For this reason, some on-line documents are accompanied by a list of keywords specified by the authors in an effort to guide the users by facilitating the filtering process.

In this way, a set of keywords is often considered a condensed version of the whole document and therefore plays an important role for document retrieval, Web page retrieval, document clustering, summarization, text mining, and so on. Since many academic journals ask the authors to provide a list of five or six keywords on the first page of an article, keywords are most familiar in the context of journal articles. However, many other types of documents could not benefit from the use of keywords, including Web pages, email messages, news reports, magazine articles, and business papers.

Although the potential benefit is large, the implementation itself is the obstacle; manually assigning keywords to all documents is a daunting task, or even impractical in that it is extremely tedious and time-consuming requiring a certain level of domain knowledge. Therefore, it is highly desirable to automate the keyword generation process. There are mainly two approaches to achieving this aim: keyword assignment approach and keyword extraction approach. Both approaches use machine learning methods and require, for training purposes, a set of documents with keywords already attached. In the former approach, there is a given set of vocabulary, and the aim is to match them to the texts. In other words, the keywords assignment approach seeks to select the words from a controlled vocabulary that best describes a document. Although this approach is domain dependent and is not easy to transfer and expand, it can generate implicit keywords

* This work was supported by the Institute of Management Research at Seoul National University.
** This work was supported by the Center for Information and Communication Business and Research at Seoul National University.
*** Ph.D. Institute of Management Research, Seoul National University, Korea.
**** Professor, Graduate School of Business, Seoul National University, Korea.
***** Director of New Services, NCSoft, Korea.
****** Corresponding author, Associate Professor, Graduate School of Business, Seoul National University, Korea.

that do not appear in a document. On the other hand, in the latter approach, the aim is to extract keywords with respect to their relevance in the text without prior vocabulary. In this approach, automatic keyword generation is treated as a classification task, and keywords are commonly extracted based on supervised learning techniques. Thus, keyword extraction algorithms classify candidate keywords in a document into positive or negative examples. Several systems such as Extractor and Kea were developed using keyword extraction approach.

Most indicative words in a document are selected as keywords for that document, and as a result, keywords extraction is limited to terms that appear in the document. Therefore, keywords extraction cannot generate implicit keywords that are not included in a document. According to the experiment results of Turney, about 64% to 90% of keywords assigned by the authors can be found in the full text of an article. Inversely, it also means that 10% to 36% of the keywords assigned by the authors do not appear in the article, which cannot be generated through keyword extraction algorithms. Our preliminary experiment result also shows that 37% of keywords assigned by the authors are not included in the full text. This is the reason why we have decided to adopt the keyword assignment approach.

In this paper, we propose a new approach for automatic keyword assignment, namely IVSM (Inverse Vector Space Model). The model is based on a vector space model, which is a conventional information retrieval model that represents documents and queries by vectors in a multidimensional space. IVSM generates an appropriate keyword set for a specific document by measuring the distance between the document and the keyword sets. The keyword assignment process of IVSM is as follows: (1) calculating the vector length of each keyword set based on each keyword weight; (2) preprocessing and parsing a target document that does not have keywords; (3) calculating the vector length of the target document based on the term frequency; (4) measuring the cosine similarity between each keyword set and the target document; and (5) generating keywords that have high similarity scores.

Two keyword generation systems were implemented applying IVSM: IVSM system for Web-based community service and stand-alone IVSM system. Firstly, the IVSM system is implemented in a community service for sharing knowledge and opinions on current trends such as fashion, movies, social problems, and health information. The stand-alone IVSM system is dedicated to generating keywords for academic papers, and, indeed, it has been tested through a number of academic papers including those published by the Korean Association of Shipping and Logistics, the Korea Research Academy of Distribution Information, the Korea Logistics Society, the Korea Logistics Research Association, and the Korea Port Economic Association. We measured the performance of IVSM by the number of matches between the IVSM-generated keywords and the author-assigned keywords. According to our experiment, the precisions of IVSM applied to Web-based community service and academic journals were 0.75 and 0.71, respectively. The performance of both systems is much better than that of baseline systems that generate keywords based on simple probability. Also, IVSM shows comparable performance to Extractor that is a representative system of keyword extraction approach developed by Turney. As electronic documents increase, we expect that IVSM proposed in this paper can be applied to many electronic documents in Web-based community and digital library.

Keywords : Information Technology, Automatic Keyword Generation, Keyword Assignment Approach, Keyword Extraction Approach, Vector Space Model, Inverse Vector Space Model, IVSM

키워드 자동 생성에 대한 새로운 접근법: 역 벡터공간모델을 이용한 키워드 할당 방법

조 원 진, 노 상 규, 윤 지 영, 박 진 수

Ⅰ. 서 론

정보의 홍수 속에서 사용자가 원하는 정보를 신속하게 얻기 위해서는 모든 자료를 검색하고 자세히 읽기보다는 가장 핵심이 되는 문서의 내용인 키워드를 통해 보다 빠른 시간에 문서의 내용을 효과적으로 이해하는 것이 보다 바람직하다. 그러나 아직까지 많은 문서들이 키워드를 포함하고 있지 않으며, 이와 같은 문서에 키워드 생성을 수작업으로 처리하는 것은 많은 시간과 비용을 요하는 작업이다. 그 결과, 문서를 대표할 수 있는 주요 키워드를 자동으로 생성하는 방법에 관한 연구들이 진행되어 왔다.

많은 양의 문서들로부터 키워드를 생성하기 위해서는 생성 방법에 대한 정확성과 효율성이 동시에 요구된다. 이를 위해 기계학습과 통계이론에 기초한 키워드 추출 중심의 키워드 생성과 관련된 연구가 진행되어 왔다[Turney, 1999; Witten *et al.*, 1999; Ercan and Cicekli, 2007]. 그러나 기존 연구의 대부분이 문서 요약(document summarization) 및 문서 범주화(document categorization)의 부분 연구로서 수행되어 왔으며, 키워드 자동 생성 시스템 자체를 위한 알고리즘 개발 및 구현을 목적으로 한 연구는 극히 소수에 불과하다. 더 나아가 국문 문서의 키워드 자동 생성에 대한 연구는 거의 찾아 볼 수 없다.

기존의 키워드 자동 생성(automatic keyword generation)을 위한 연구는 크게 다음의 두 가지 접근법으로 구분할 수 있다[Witten *et al.*, 1999]. 첫째는 키워드 자동 추출(automatic keyword extraction)로서 이는 문서에 포함된 용어들의 정보에 기초하여 '문서 안에 포함된 용어들 중'에서 키워드를 생성하는 것이다. 다른 하나는 키워드 자동 할당(automatic keyword assignment)의 방법으로서 키워드를 생성하고자 하는 문서를 분석하여 그것과 관련 있는 키워드를 '미리 정의된 사전으로부터' 키워드를 생성하는 방법이다. 키워드 자동 생성과 관련된 연구들의 대부분이 키워드 자동 추출의 방법을 적용하고 있다. 이는 키워드 자동 추출의 방식이 따로 사전을 구축할 필요가 없으며, 문서에 포함된 용어들의 가중치를 통계적으로 결정하여 비교적 간단한 방식으로 키워드를 생성해 낼 수 있기 때문이다. 그러나 이 방법은 문서에 표현된 용어로 키워드가 생성되기 때문에 잘못된 어휘나 표현으로 키워드가 지정될 수 있는 문제점이 있다. 또한 문서를 구성하고 있는 용어들이 단일어 중심으로 분석되기 때문에 복합어로 된 키워드를 생성해 내기 위해서는 추가적인 노력이 필요하다. 반면, 키워드 자동 할당의 방법은 사전을 구축하는데 어려움이 있고, 지속적으로 사전을 관리하는데 어려움이 있다. 그러나, 오늘날 다양한 분야에서 어휘 사전 온톨로지가 구축됨에 따라 별도의 사전 구축에 필요한 노력이 줄어들게 되었다. 대표적인 국문 어휘 사전으로, 한국어 워드넷이라고 할 수 있는 KorLex[Yoon *et al.*, 2009]가 공개되었다. 더 나아가 사전에는 단일어, 복합어의 구분 없이 다양한 형태의 어휘를 포함하고 있기 때문에 복합어로 된 키워드를 생성해 내는데 큰 추가적인 노력이 필요 없다.

본 연구에서는 입력된 질의어에 대해서 가장 유사한 문서를 찾아주는 검색분야의 대표적인

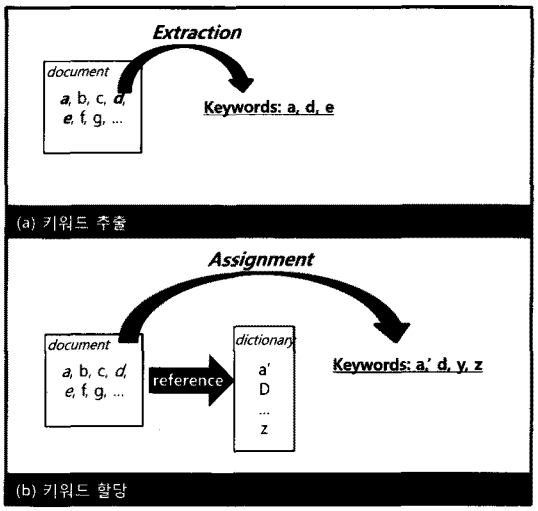
알고리즘인 벡터공간모델(Vector Space Model)을 역으로 적용하여 키워드를 자동으로 생성하는 역 벡터공간모델(Inverse Vector Space Model: 이하 IVSM)을 제안한다. 이 모델은 키워드 자동 할당의 접근법으로서 미리 잘 구축된 사전에 해당되는 키워드 집합들(keyword sets)이 필요하며, 이러한 키워드 집합들에 기초하여 입력 문서에 가장 적합하다고 판단된 키워드를 생성한다. 특히, 제안하는 모델은 문서 범주화나 요약의 부분 모델이 아닌 전적으로 키워드 생성 시스템 자체를 위한 것으로서 자연어로 처리된 글을 실시간으로 분석하여 키워드를 자동 생성하므로 문서를 공유하는 다양한 웹 기반 서비스 등에 직접적으로 적용 가능하다.

본 연구의 구성은 다음과 같다. 제 Ⅱ장에서는 관련 연구로서 키워드 자동 생성에 관한 연구들을 소개한다. 이어서 제 Ⅲ장은 본 연구에서 제안하는 IVSM에 대해 구체적으로 소개하고, 제 Ⅳ장은 IVSM에 대한 평가결과를 제시한다. 마지막으로 제 Ⅴ장은 본 연구를 종합하며 마무리한다.

Ⅱ. 관련 연구

2.1 키워드 자동 생성(Automatic Keyword Generation)

키워드 자동 생성과 관련된 연구는 크게 키워드 추출 접근법(keyword extraction approach)과 키워드 할당 접근법(keyword assignment approach)의 두 가지 방법으로 구분할 수 있다(<그림 1> 참조). 키워드 추출 방법의 경우 문서 집합 내에서 키워드를 추출하는 것을 기본으로 한다. 반면 키워드 할당 접근법은 정제된 어휘사전을 이용하여 문서에 적합할 것으로 판단되는 용어를 키워드로 할당하는 기법이다.



<그림 1> 키워드 자동 추출 vs. 키워드 자동 할당

2.1.1 키워드 추출 접근법(Keyword Extraction Approach)

기존의 키워드 추출 접근법(keyword extraction approach)은 단순 통계적 접근법과 기계학습에 기반한 접근법으로 구분할 수 있다. 통계이론에 기반하여 키워드를 추출하는 경우는 *tfidf* (Term Frequency Inverse Document Frequency)에 기반한 연구가 주종을 이루고 있다. Lee and Bae[2002]는 앵커 텍스트(Anchor Text)의 용어들이 키워드로 적합한지를 *tfidf*를 이용하여 검증하였다.

tfidf 이외에도 주성분 분석, 공기정보(co-occurrence), 확률적 그래프 모형 등 다양한 통계적 기법들이 키워드 추출을 위해 사용되기도 한다. Shin *et al.*[2000]은 텍스트 문서의 키워드를 추출하고 문서를 주제별로 분류하기 위해 확률적 그래프 모델을 사용하는 방법을 제안했다. 텍스트 문서 데이터를 문서와 용어의 쌍으로 표현하여 확률적 그래프 모델을 학습시켰으며, 이를 위해 정의된 우도(likelihood)를 최대화하기 위한 EM (expected maximization) 알고리즘을 사용하였다. Lee *et al.*[2002]은 문서의 내용을 대표할 수

있는 키워드를 추출하는데 있어 다변량 통계 분석 기법 중의 하나인 주성분 분석을 이용하는 모델을 제안했다. 이 모델은 고유 값과 고유벡터를 이용하여 문서자체 내에서 용어의 흐름을 정량화하고, 그 정보를 이용하여 문서 자체 내에서의 발생 빈도와 공기정보를 이용하여 주제어를 추출하였다. Matsuo and Ishizuka[2004]는 여러 문서로 구성된 코퍼스(corpus)가 아닌 하나의 문서에 적용할 수 있는 새로운 키워드 추출 알고리즘을 제시했다. 우선 빈발 용어가 추출되고, 그 다음은 각 용어와 그 빈발 용어가 같은 문장에 존재하는 공기정보가 계산된다. 만약 용어 a와 그 빈발 용어들 사이의 공기 확률 분포가 빈발 용어들의 특정 하위집합에 편의 되면, 용어 a는 키워드가 될 수 있는 것으로 판단한다. 분포의 편의 정도는 카이제곱에 의해 평가된다. 그들이 제시한 알고리즘은 코퍼스를 사용하지 않고도 *tfidf*를 적용한 결과와 유사한 성능을 보였다.

기계학습 기법을 적용하여 키워드를 추출하는 문제는 일종의 분류(classification) 문제와 동일하게 간주된다. 즉, 특정 문서에 포함되어 있는 용어가 키워드인지 아닌지의 이진분류 문제와 동일하게 간주하여 문제를 해결하고자 한다. Turney [1999, 2000]는 다섯 개의 분야로 나눠지는 총 652개의 문서들에 대해 C4.5 의사결정나무(decision tree)를 적용하여 키워드 추출을 위한 학습 모델을 만들고 평가했으며, 실험결과 75%의 성공률을 보였다. 보다 높은 정확도를 달성하기 위해 유전 알고리즘(genetic algorithm)을 함께 적용한 하이브리드 모델인 시스템을 개발했으며, 이것이 GenEx라는 키워드 추출 시스템이다. GenEx에서 유전 알고리즘은 규칙들의 파라미터들을 조정하여 훈련 문서들에서 올바르게 식별된 키워드들의 수를 최적화하는 역할을 한다. 이후 GenEx는 Extractor라는 소프트웨어로 공개되었다.

GenEx와 함께 키워드를 추출하는 대표적인 시스템인 Kea[Frank *et al.*, 1999; Witten *et al.*, 1999]는 뉴질랜드 Waikato 대학에서 개발된 시스템이다. Kea(an algorithm for automatically extraction keyphrases from text)는 우선 다른 키워드 추출 기법과 유사한 방법으로 후보 키워드들을 추출한다. 후보 키워드로 추출되기 위해서는 용어들 사이에 불용어가 있어서는 안 된다는 조건만 만족하면 된다. 다음 단계로 Kea는 이미 사람들에 의해 키워드가 할당된 문서집합을 이용하여 학습을 통해 모델을 구축한다. 이때 학습 알고리즘으로 나이브 베이지안 기계학습 알고리즘을 사용하였으며, 모델 구축을 위해 사용한 특징들로 *tfidf* 가중치와 거리(distance)를 사용하였다. 여기서 거리는 해당 키워드가 문서의 어느 부분에서 처음으로 나타났는가를 의미하는 것으로 보다 일찍 출현한 용어일수록 그 문서를 표현함에 있어 영향력이 클 것으로 기대한다. 이 시스템은 현재 뉴질랜드 디지털 도서관에 구현되어 있으며 다른 키워드 추출 시스템에 비하여 상당히 효율적인 시스템으로 평가 받고 있다.

Ercan and Cicekli[2007]는 키워드 추출을 위해 어휘 사슬(lexical chain)을 적용하였다. 어휘 사슬은 문서 요약을 위해 광범위하게 사용되어왔지만, 키워드 추출 문제에의 적용은 이들 연구가 처음이다. 어휘 사슬을 통해 텍스트의 어휘적 결속 구조(lexical cohesive structure)를 포착할 수 있다. 이와 같은 어휘 사슬 구축을 위해서는 용어 간의 관계를 명확히 정의해 놓은 데이터베이스가 필요하며, Ercan and Cicekli는 어휘 사슬 구축 시 이러한 데이터베이스로서 워드넷(WordNet)을 이용하였다. 구축된 어휘 사슬을 통해 각 용어의 어휘 사슬 점수가 계산되는데, 이것은 어휘 사슬에 존재하는 관계들, 즉 유의어, 동의어, 상위어, 하위어 같은 항목들에 기초하여 계산된다. 어휘 사슬을 통해 각 용어의 어휘 사슬 점수(lexical chain score of a word), 각 용어의 직접적인 어휘 사슬 점수(direct lexical chain score of a word), 각 용어의 어휘 사슬 범위 점수(lexical chain span score of a word), 그리고 각 용어의 직접적인 어

휘 사슬 범위 점수(direct lexical chain span score of a word)의 네 가지 특징 벡터들이 도출됐다. 제안하는 새로운 시스템의 평가를 위해, 키워드 추출 시스템에서 기본적인 특징으로 많이 사용되는 세 가지 특징 벡터인 첫 등장 위치(first occurrence position), 용어 빈도(word frequency), 그리고 마지막 등장 위치(last occurrence position)를 사용한 기본시스템(baseline system)이 구축되었다. 여기에 어휘 사슬을 통해 추출된 특징들 네 가지를 추가하여 총 일곱 가지의 특징 벡터를 사용한 추출 시스템을 개발했다. 그 결과 세 가지 특징 벡터를 사용한 기본시스템은 5개의 키워드를 추출하는 경우 17%의 정확도를 보였으나, 어휘 사슬을 통해 도출된 특징들을 추가하여 일곱 개의 특징 벡터를 사용한 시스템은 5개의 키워드를 추출하는 경우 45%로 기본시스템보다 더 높은 정확도를 보였다.

2.1.2 키워드 할당 접근법(Keyword Assignment Approach)

키워드 할당 방법은 기본적으로 모든 잠재적 키워드들이 미리 정의된 어휘 목록 혹은 사전에 포함되어 있다고 가정한다. 따라서 키워드 할당 방법은 미리 정의된 사전을 계속적으로 관리해야 하는 어려움이 있으며, 특히 신조어에 대한 지속적인 업데이트를 필요로 한다[Frank *et al.*, 1999]. 이와 같은 어려움으로 인하여, 키워드 할당 방법에 대한 연구는 키워드 추출 방법을 적용한 키워드 생성 연구에 비해서 그 수가 매우 적으며, 수행된 연구의 대부분 역시 문서 범주화를 위한 부분 연구로서 수행되었다. 문서 범주화는 문서를 미리 정의된 카테고리로 나누는 것으로, 이 역시 분류 문제로 인식되어 이 문제를 다루는 기계학습 기법이 문서 범주화에 적용되었다.

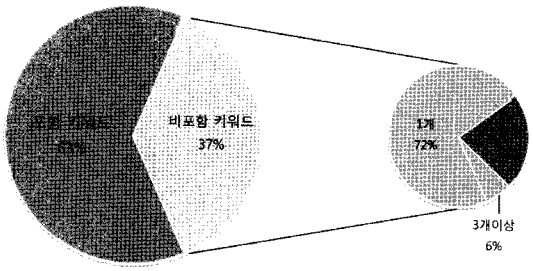
비록 키워드 할당 방식이 사전의 유지 및 관리에 대한 어려움이 있으나, 사전이 잘 관리되고 유지된다면 키워드로 할당되는 어휘들이 정제될 수 있고, 그 결과 통일성 있고 일관된 표현으로 키워드들이 축적될 수 있다는 장점이 있다. 예를 들어, "DEA"란 용어는 국문으로 "자료 포락 분석", "자료 포괄 분석" 등 다양한 표현으로 사용되고 있는데, 만약 사전에 "자료 포락 분석"으로 정제되어있다면 문서에 "자료 포괄 분석"으로 표현이 되어 있더라도 할당되는 키워드는 "자료 포락 분석"으로 지정한다. 따라서 DEA 관련 연구들의 국문 키워드 부분에 DEA의 명칭으로 산발적인 표현이 사용되는 것이 아닌 "자료 포락 분석"으로 일관되게 키워드가 지정된다.

키워드 할당과 관련된 연구는 기계학습 기법을 주로 이용하고 있다. 즉, 기계학습 기법을 이용하여 특정 문서에 대해서 여러 카테고리 중에 가장 관련된 카테고리들로 분류하여 그 카테고리 안에 포함된 용어로 키워드를 생성하는 접근법을 취하고 있다. Zhang and Xu[2009]는 citation-kNN 방법에 기초하여 자동으로 키워드를 할당하는 방법을 제안했다. 문서 범주화 분야에서 좋은 성능을 보여주는 메모리 기반 학습 기법인 kNN(k-Nearest Neighborhood)을 확장한 citation-kNN 기법을 적용하여 문서에 포함되지 않은 용어들까지도 키워드로 할당하는 알고리즘을 제안했다. 실험결과 정확도(precision)와 재현율(recall)이 향상되었으며, 문서에 포함되지 않은 내재적인 어휘들도 키워드로 도출할 수 있었다.

Ⅲ. 역 벡터공간모델을 이용한 키워드 할당 방법

앞에서 살펴본 바와 같이, 키워드를 자동으로 생성하는 방법에는 크게 글 내용에 포함된 용어의 가중치를 분석해서 그 가중치의 크기에 따라 키워드를 선별하는 "키워드 추출" 방식과 글 내용을 분석해서 글과 가장 관련된 키워드를 미리 정의된 사전으로부터 선택해서 생성하는 "키워드 할당"의 두 가지 접근법으로 구분된다. Turney [1997]의 연구에 따르면, 키워드들의 약 65%~

90%는 논문의 전문에서 발견된다고 한다. 즉, 키워드로서 지정된 용어들이 논문상에 포함되지 않을 가능성도 10%~35% 정도 있다는 것을 시사한다.



<그림 2> 지정된 키워드의 본문 포함 여부에 대한 조사결과

실제로 국문의 문서에도 키워드 할당 방법 적용의 필요성에 대해 사전 실험을 수행했다. 이를 위해 한국항만경제학회지 논문 57편, 대한경영교육저널 논문 30편에 대해서 저자가 지정한 키워드가 실제로 본문에 포함되었는지 여부를 조사해보았다.[1] 그 결과 <그림 2>에서 보이는 것과 같이, 총 87편의 논문 중 37%에 해당되는 32편의 논문들이 논문에 포함되지 않은 용어나 문구로 키워드가 지정되어 있었다. 이는 본문에 포함된 용어들에서 키워드로 생성하는 키워드 추출 방법으로 생성될 수 없는 것들로, 키워드 추출 방법 적용에 있어 한계가 있음을 보여준다. 이에 본 연구는 키워드 추출이 아닌 키워드 할당에 대한 연구로서 검색 분야에서 적용되어온 벡터공간모델을 역으로 적용하여 미리 정의된 키워드 집합들로부터 키워드를 생성해내는 새로운 알고리즘인 역 벡터공간모델(Inverse Vector Space Model)을 제안한다. 이어서, 제안하는 역 벡터공간모델에 기본 아이디어를 제공한 벡터공간모델에 대한 간단한 소개와 함께 역 벡터공간모델에 대해 구체적으로 소개한다.

1) 본 사전 실험은 실험의 편의를 위해 원본이 '텍스트 pdf' 형식의 파일인 논문들에 대해서 수행되었다. 즉, 이미지 스캐닝을 통해 만들어진 '이미지 pdf' 형식은 단어의 문서 내 검색이 불가능한 반면, 텍스트 pdf 형식의 파일은 '찾기 기능'을 이용해 단어의 문서 내 검색이 가능하여 지정한 키워드가 실제로 논문에 포함되었는지를 조사하는데 용이하다.

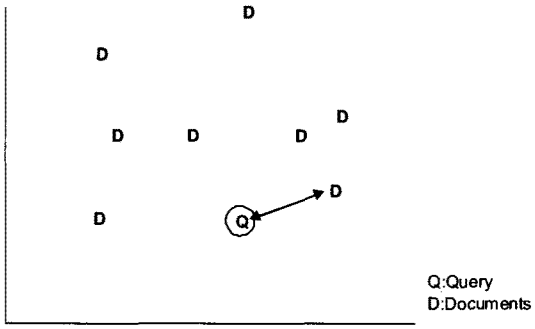
3.1 벡터공간모델(Vector Space Model)

최근에는 많은 정보를 손쉽게 접할 수 있는 검색 서비스를 제공하는 사이트들이 많아지고 있으며, 원하는 정보를 보다 빠르고 쉽게 찾기 위한 사용자들의 이용 또한 나날이 증가하고 있다. 검색 서비스를 제공하는 사이트에서는 통상적으로 인터넷 상에 존재하는 원시데이터를 수집하여 해당 자료 별로 키워드를 선정한 후 데이터베이스로 구축하고, 사용자들이 찾고자 하는 자료의 일부 키워드를 입력하여 검색을 요청하면 해당 키워드로 지정되어 있는 자료들을 사용자들에게 제공하는 형태로 서비스를 수행하고 있다.

이때 검색 서비스를 제공하는 사이트에서는 사용자들의 검색에 따라 제공되는 자료를 문서의 정확도, 중요도 등에 따라 우선순위를 가지는 문서들을 상위에 배치하여 사용자들에게 제공한다. 이와 같은 문서의 중요도를 분석하는 많은 방법 중에서 벡터공간모델(Vector Space Model)은 정보 필터링, 문서 내에서의 정보검색, 색인과 유사도를 계산하기 위한 수학모델로서, 다차원 선형공간에서의 벡터 정보를 이용하여 자연어를 포함한 문서의 중요도를 분석하기 위한 방법을 제시하고 있다[Yang and Huh, 2004].

"Term Vector Model"이라고도 불리는 벡터공간모델은 코넬 대학교의 Gerald Salton에 의해 개발되었다. SMART(System for the Mechanical Analysis and Retrieval of Text)에 벡터공간모델이 적용되어 개발된 이래, 현재 내용기반 웹 정보검색에서 가장 전형적인 검색 모델로 간주되고 있다[Salton and McGrill, 1983]. 벡터공간모델에서 각 문서는 그 문서가 포함하고 있는 색인용어

의 벡터로 나타내지며, <그림 3>처럼 문서의 유사도는 벡터에 위치한 용어들 간의 거리로 계산해낼 수 있다 라는 것이 이 모델의 대전제이다.



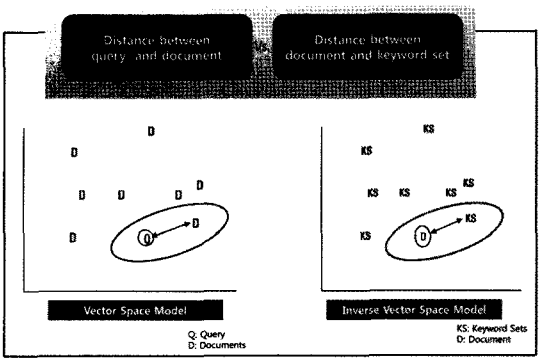
<그림 3> 벡터공간모델의 유사도 계산

구체적으로 살펴보면, 벡터공간모델에서 용어와 문서는 *k*차원 공간의 벡터로 인코딩된다[Berry *et al.*, 1999]. *k*는 독립적인 용어와 개념 또는 텍스트와 관련된 클래스의 수에 따라 결정된다. 그러므로 각각의 벡터 성분은 대응되는 용어, 개념 및 클래스의 중요성을 반영한다. 벡터공간모델 상에서는 각 문서들과 사용자 질의는 *k*차원 공간 속의 벡터들로 취급되며, 이때 각 차원들은 색인용어들로 표현된다.

벡터공간모델을 사용하기 위해서는 문서의 벡터 공간에 있는 용어의 가중치(weights)를 계산하고 있어야 한다. 이를 위해서 앞의 제 II장에서 소개된 *tfidf*가 주로 사용되고 있다. 벡터공간모델의 주요 장점은 용어의 가중치 기법이 검색 성능을 향상시키고, 질의 조건에 근접한 문서 검색이 가능하며, 거리 측정 기법 즉 코사인 순위화 등의 기법이 문서들을 질의어를 기준으로 정렬해 줄 수 있다는 점이다. 반면 문서가 포함하고 있는 주요 색인어가 상호 독립적이어야 한다는 가정이 단점으로 지적된다. 이를 극복하기 위해 온톨로지를 적용하여 주요 색인어를 분류하고 거리를 고려하는 방법도 시도되고 있다.

3.2 역 벡터공간모델(Inverse Vector Space Model: IVSM)

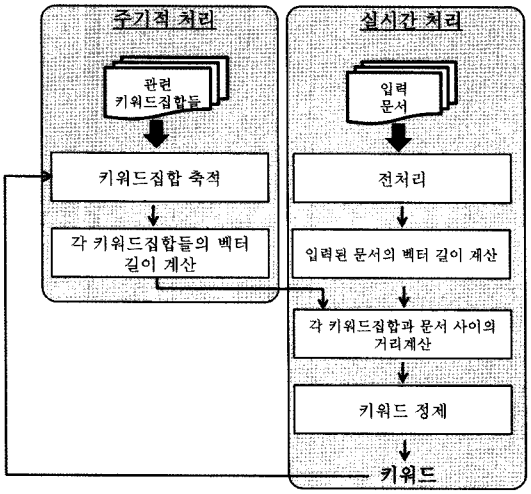
앞서 설명한 벡터공간모델은 사용자가 질의어를 입력한 경우, 그 질의어와 가장 유사한 문서를 찾아주기 위해, 질의어와 각 문서들의 거리를 측정한다. 본 연구는 벡터공간모델을 역으로 적용하여 특정 문서에 적합한 키워드를 할당하는 새로운 아이디어를 제안한다. 즉, 기존의 검색분야에서 벡터공간모델은 많은 문서들 중에 입력한 키워드와 가장 근접한 문서를 찾는데 적용되어 왔으나, 본 연구에서는 여러 키워드 중에서 특정 문서와 가장 유사한 키워드 집합을 찾아내기 위해서 역 벡터공간모델(Inverse Vector Space Model)을 적용하는 새로운 방법을 제시한다(<그림 4> 참고). 그 결과 문서에 입력된 용어들과 키워드들을 비교하여 각 문서에 포함된 용어들에 기반하여 적합한 키워드를 추천할 수 있다.



<그림 4> 벡터공간모델과 역 벡터공간모델(IVSM)의 비교

IVSM의 키워드를 할당하는 과정은 <그림 5>와 같다. IVSM은 실시간으로 처리되는 모듈과 주기적으로 처리되는 모듈로 구분될 수 있다. 키워드를 할당하고자 하는 목표 입력 문서를 분석하고 가장 유사한 키워드 집합을 선택하여 할당하는 과정은 실시간으로 처리된다. 반면, 키워드 집합은 주기적으로 수집되어 각 키워드 집합들

의 벡터 값이 계산된다. 각각의 단계에 대해서 구체적으로 소개하면 다음과 같다.



<그림 5> 역 벡터공간모델 키워드 할당 프로세스

3.2.1 키워드 집합 벡터화

IVSM은 키워드 할당 모델로서, 기본적으로 일종의 사전과 같은 수많은 키워드 집합이 필요하다. 키워드 집합(keyword set)이란 한 문서를 대표하는 키워드들의 모임을 의미한다. 예를 들어 문서 A에 키워드로 "시맨틱웹", "온톨로지", "웹 3.0"이 지정되어 있는 경우, {시맨틱웹, 온톨로지, 웹 3.0}이문서 A를 대표하는 하나의 키워드 집합이 된다. 이와 같은 각각의 키워드 집합이 축적되어 전체 키워드 집합을 형성하게 된다. 필요한 키워드 집합에 대한 수집이 완료되면, 각각의 키워드 집합에 대해 벡터 길이가 계산되어야 한다. 키워드 집합 벡터화를 위해서는 키워드 집합에 포함된 각각의 키워드들의 가중치에 대한 정보가 필요하다. 이를 위해 다양한 가중치 부여 기법이 적용될 수 있으나 본 연구에서는 기존 *tf*(term frequency) 가중치에서 변경된 수정*tf* 가중치(augmented term frequency weight: 이하 *kw*)로 각 키워드의 가중치를 부여했다. 이것은 키워드 집합에서 가장 많이 등장하는 키워드에 대한 상대적인 빈도수로서 가중치를 설정하는 방식이다. 각 키워드들의 가중치를 바탕으로 다음의 식 (1)과 같이 모든 키워드 집합에 대해서 벡터 길이가 계산된다. 식 (1)에서, KF는 키워드 할당을 위해 수집된 키워드 집합 전체에서 각 키워드가 몇 번 등장하는지를 의미한다. 예를 들어, 경영정보와 관련된 논문들에서 축적된 전체 키워드 집합의 개수가 1000개이고, "온톨로지"라는 키워드가 20개의 키워드 집합에 포함되어 있다면, KF는 20/1000의 값을 갖게 된다.

$$|KS_i| = \sqrt{\sum_j kw_i^2} \quad (1)$$

kw(keyword weight) = 0.5 + 0.5 * (KF/Max KF)

KF = 키워드 집합 전체에서 키워드가 포함된 빈도수

Max KF = 키워드 집합 전체에서가장 빈도수가 높은 키워드의 빈도수

3.2.2 문서 벡터화

키워드 생성을 하고자 하는 목표 문서(target document)는 그 문서에 포함되어 있는 용어들을 기반으로 벡터화 된다. 문서를 벡터화 할 때 적용하는 용어 가중치는 여러 가지가 있을 수 있다. 예를 들어, 초기에는 정보검색 분야의 대표적인 가중치 기법으로 문헌과 문헌집합 내 출현 정보를 조합한 *tfidf* 가중치 기법을 주로 사용해 왔으나, 최근에 와서 범주 정보의 사용을 적극적으로 검토하는 등 다양한 시도가 진행되고 있다[Kim, 2008]. 본 연구에서는 문서 내 출현 빈도에 기초한 빈도 가중치와 문서 내 포함된 그 용어의 키워드 집합 내에서의 가중치를 함께 고려하여 가중치를 설정했다. 우선, 빈도 가중치(frequency weight: *fw*)는 형태소 분석 결과 얻어진 용어의

문서 내 출현 빈도에 기초한 가중치로, 이것은 문서에 포함된 전체 용어 수에 대한 상대적인 비율로 계산된다. 이것에 미리 정의된 키워드 집합에서 가지는 중요도에 대한 가중치를 반영하기 위해서 키워드 가중치(keyword weight: *kw*)를 곱하여 문서에 포함된 각 용어의 가중치가 계산된다. 그 결과 최종적으로 목표 문서는 다음의 식 (2)에 의해 벡터 길이가 계산된다. 이때, IVSM은 키워드 집합에 존재하는 키워드를 바탕으로 키워드를 할당하는 모델이므로, 문서에 포함되어 있지만 축적된 키워드 집합에 포함되지 않은 용어는 문서 벡터화에 포함시키지 않는다.

$$|D| = \sqrt{\Sigma_i (fw_i * kw_i)^2} \quad (2)$$

fw_i = 문서 내 i번째 용어의 출현 빈도에 기초된 가중치

kw_i = i번째 용어가 키워드 집합에서 가지는 가중치

3.2.3 키워드 집합과 문서간 유사도 계산

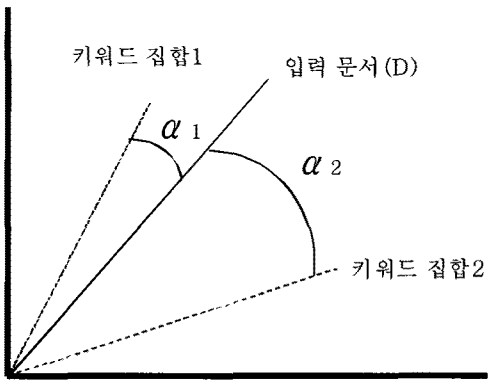
각 키워드 집합과 목표 문서와의 유사도가 계산된다. 유클리디안 거리, 자카드 계수 등 유사도를 판단하는데 적용할 수 있는 다양한 척도가 있으나 본 연구에서는 코사인 유사도 공식을 이용해 유사도를 계산한다(식 (3) 참고). IVSM은 벡터 공간 상에서 키워드 집합과 목표 문서 사이의 유사도를 계산해야 하므로, 단순히 2차원 좌표 상의 거리를 측정하는 유사도 공식이 아닌, 벡터의 성질을 이용한 유사도 측정 방법인 코사인 유사도 공식을 이용하는 것이 적합하다고 판단된다. 더욱이 코사인 유사도 공식은 벡터 값을 갖는 문서의 유사도를 판단하는데 있어 탁월한 성능을 보이는 척도로 평가 받고 있다[Wan, 2007].

$$\cos\theta = \frac{\Sigma_i fw_i * kw_i}{|KS_i| * |D|} \quad (3)$$

fw_i = 문서 내 i번째 용어의 출현 빈도에 기초된 가중치

kw_i = i번째 용어가 키워드 집합에서 가지는 가중치

코사인 유사도 공식에서 의 의미는 문서 벡터 값과 키워드 집합 벡터 값 사이의 각도를 의미한다. 즉, θ가 0이면, 문서와 키워드 집합은 벡터 상에 가장 가까이 위치하며, 두 벡터 간의 유사도는 가장 큰 값인 1의 값을 갖게 된다. 예를 들어, 다음의 <그림 6>에서 보면, 입력된 문서(D)는 키워드 집합1에 더 가까이 위치해 있으므로 키워드 집합1이 입력 문서에 대해 키워드 집합2보다 더 유사한 키워드 집합이라고 할 수 있다.



<그림 6> 입력 문서와 키워드 집합 간 유사도 비교

Ⅳ. 시스템 평가

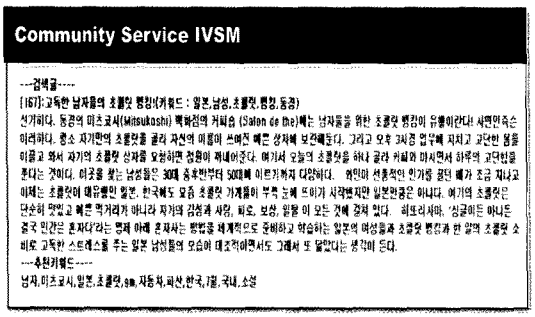
4.1 역 벡터공간모델 기반 키워드 할당 시스템 구현

다양한 도메인에서의 적용 가능성을 평가하기 위해서, IVSM을 두 개의 시스템으로 구현했다. 우선, 다양한 주제에 대해 비교적 짧은 글로 사용자간 개인의 의견을 교환하는 웹 기반 커뮤니티

서비스에 IVSM을 적용했다. 또한, 독립 시스템(stand-alone IVSM)으로서 키워드 집합을 축적하고 문서를 입력하여 키워드를 생성하는 독립 시스템을 구현했다.

4.1.1 웹 기반 커뮤니티 서비스 시스템 구현

다양한 사용자층을 포함하며 다양한 토픽에 대해 사용자간 개인의 의견을 교환할 수 있는 웹 기반 커뮤니티 서비스에 IVSM의 적용 가능성 및 정확도를 분석하기 위해 현재 서비스 되고 있는 커뮤니티 서비스에 IVSM을 구현했다(<그림 7> 참조). IVSM이 적용된 웹 커뮤니티 서비스는 패션, 영화, 사회적 이슈, 건강 등과 같은 그 시대의



<그림 7> 웹 기반 커뮤니티 서비스에 구현된 IVSM

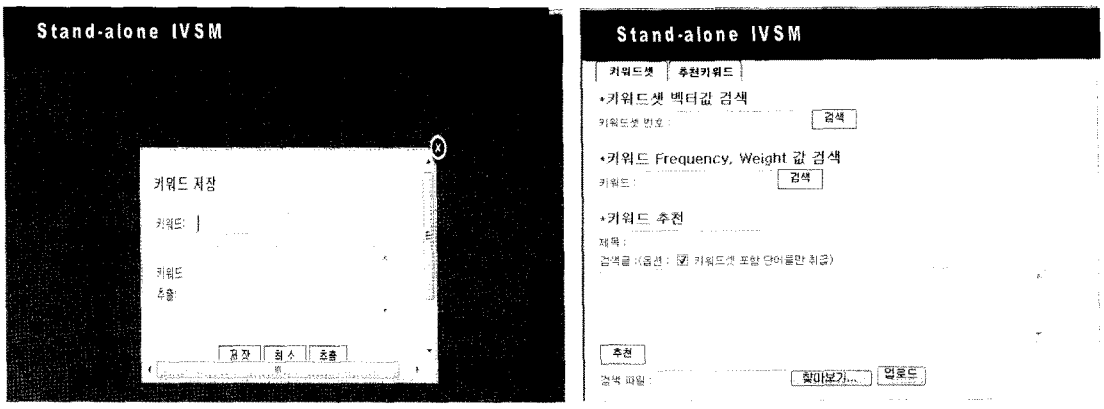
트렌드에 대한 의견과 지식을 공유하는 실제 서비스되고 있는 커뮤니티 서비스이다. 구현된 IVSM에 사용된 키워드 집합으로는 적용된 웹 기반 커뮤니티 서비스에 기존에 글쓴이들이 직접 지정한 수많은 키워드 집합을 이용했다. 이 키워드 집합들은 개인 사용자들이 작성한 글에 스스로 입력한 키워드들로 구성되었으며, 실험 당시 약 2,000개 이상의 키워드 집합이 사용되었다.

4.1.2 독립 IVSM 시스템 구현

앞서 구현된 웹 기반 커뮤니티 서비스의 글들은 다양한 유저의 다양한 관심사에 대한 글들이 입력되므로, 하나의 도메인으로 집중되는 경향은 거의 없다고 할 수 있다. 따라서 특정 도메인에 한정한 경우에 대한 시스템 성능을 분석하기 위해서 별도의 시스템을 구현할 필요가 있다. 이 시스템은 키워드 집합 입력 부분과, 이를 바탕으로 키워드를 생성하는 추천키워드 부분으로 구성된다(<그림 8> 참조).

4.2 IVSM 성능 평가

IVSM의 성능의 평가는 문서 작성자가 지정한 키워드와 IVSM에 의해 생성된 키워드가 얼마나



<그림 8> 독립 IVSM 시스템의 키워드 집합 입력 부분과 키워드 추천 부분

일치되는지를 정확도와 재현율, 그리고 F-measure에 기반하여 평가했다.[2] 정확도와 재현율, F-measure의 계산 방식은 다음과 같다.

$$\text{정확도(precision)} = \frac{\text{글쓴이에 의해 지정된 키워드} \cap \text{IVSM에 의해 생성된 키워드}}{\text{IVSM에 의해 생성된 키워드}} \quad (4)$$

$$\text{재현율(recall)} = \frac{\text{글쓴이에 의해 지정된 키워드} \cap \text{IVSM에 의해 생성된 키워드}}{\text{글쓴이에 의해 지정된 키워드}} \quad (5)$$

$$\text{F-measure} = 2 \times \frac{\text{precision} \times \text{recal}}{\text{precision} + \text{recal}} \quad (6)$$

평가는 (1)축적된 키워드 집합의 개수에 대한 의존도, (2)입력 문서의 크기, (3)다양한 분야의 적용 가능성, 그리고 (4)영문 문서를 바탕으로 한 기존 시스템(Extractor)과의 성능 비교의 네 가지 측면에 대해서 수행되었다. 앞 절에 설명된 바와 같이, IVSM은 '커뮤니티 서비스 키워드 생성 시스템'과 '독립 시스템'의 두 가지 시스템으로 구현되었다. 입력 문서의 크기와, 축적된 키워드 집합의 개수에 대한 의존도의 평가, 영문 문서를 바탕으로 한 Extractor와의 성능 비교에는 독립시스템이 사용되었으며, 다양한 분야의 적용 가능성에 대한 평가에는 두 가지 시스템이 모두 사용되었다. 특히, IVSM을 통해 생성된 키워드의 정확도가 우수한지에 대해서 객관적으로 평가하기 위해 기본시스템(baseline system)과 비교 평가를 수행했다. 기본시스템은 확률 기반인 용어빈도(term frequency)에 기초하여 키워드를 생성해내는 시스템으로 문서 출현 가능성이 높은 상위 빈도수를 갖는 용어를 키워드로 생성해내는 시스템이다.

4.2.1 키워드 집합의 개수에 대한 의존도 평가

본 연구에서는 축적되는 키워드 집합의 개수에 대한 의존 정도를 보기 위해서 키워드 집합의 개수를 다르게 하여 IVSM의 성능을 평가했다. 이를 위해, 유통 및 물류라는 특정 도메인을 정해 관련 키워드 집합을 수집했다. 유통 및 물류 분야의 관련된 모든 키워드 집합을 수집하기 위해서 우리말 키워드를 제공하는 모든 관련 저널을 조사했으며, 그 결과 한국유통경영학회지, 한국로지스틱스학회지, 한국물류학회지, 한국항만경제학회지, 한국해운물류학회지가 선정되었다. 이 저널들로부터 키워드 집합을 수집했으며, 400개, 450개, 500개, 600개, 700개 그리고 830개로 키워드 집합의 수를 변화하면서 이에 따른 성능을 평가해보았다. 특히, 400개와 450개의 키워드 집합 축적의 경우 총 830개의 키워드 집합으로부터 400개와 450개의 키워드 집합을 10번을 무작위로 추출하여 각각을 축적하고 이를 평가하는 실험을 각각 10회에 걸쳐 진행하였다. 또한 500개와 600개의 축적실험의 경우는 5번을 무작위로 추출하여 이를 축적하여 실험을 진행하였고, 700개는 2회의 무작위 추출을 통해 실험을 수행하였다.

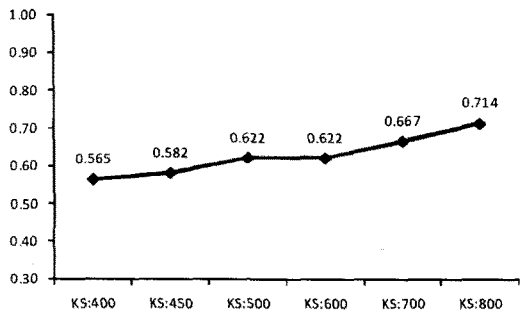
평가 데이터(evaluation data)는 키워드 집합을 제공한 5개 저널의 최근 논문들 중 30개를 선정했으며, 초록이 사용되었다. 단, 평가 데이터로 사용된 논문들의 키워드는 830개의 키워드 집합에 포함시키지 않았다. 평가 데이터로 사용된 논문들에 저자가 할당한 키워드의 개수는 평균적으로 4개였으며, 이에 IVSM을 통해 생성되는 키워드 역시 4개로 제한하였다. 각 키워드 집합의 개수를 변화시키며 실험한 결과의 정확도와 재

2) 문서작성자가 표기한 단어와 시스템이 생성한 단어 간에 비록 표기는 다를지라도 핵심어 혹은 핵심 의미가 같은 경우 일치하는 것으로 판단했다. 예를 들어, 문서작성자가 해당 문서의 키워드로 "MIS"를 지정했고, IVSM은 "경영정보시스템"이란 단어를 생성한 경우, "MIS"와 "경영정보시스템"은 통상적으로 같은 내용을 의미하는 바, IVSM이 올바른 키워드를 생성한 것으로 평가했다.

<표 1> 키워드 집합 개수 별 정확도 및 재현율 평가 결과

D#	KS#: 400		KS#: 450		KS#: 500		KS#: 600		KS#: 700		KS#: 830	
	정확도	재현율	정확도	재현율	정확도	재현율	정확도	재현율	정확도	재현율	정확도	재현율
1	0.43	0.87	0.65	0.87	0.55	0.73	0.70	0.93	0.63	0.83	0.75	1.00
2	0.85	0.85	0.83	0.83	0.65	0.65	0.90	0.90	0.88	0.88	1.00	1.00
3	0.63	0.63	0.65	0.65	0.55	0.55	0.70	0.70	0.75	0.75	0.75	0.75
4	0.50	0.67	0.53	0.70	0.65	0.87	0.50	0.67	0.63	0.83	0.50	0.67
5	0.60	0.60	0.63	0.63	0.65	0.65	0.65	0.65	0.63	0.63	0.75	0.75
6	0.50	0.50	0.58	0.58	0.60	0.60	0.60	0.60	0.63	0.63	0.75	0.75
7	0.45	0.60	0.43	0.57	0.65	0.87	0.35	0.47	0.38	0.50	0.50	0.67
8	0.88	0.58	0.83	0.55	0.55	0.37	0.95	0.63	0.63	0.42	1.00	0.67
9	0.58	0.77	0.60	0.80	0.65	0.87	0.75	1.00	0.63	0.83	0.50	0.67
10	0.40	0.53	0.43	0.57	0.75	1.00	0.45	0.60	0.63	0.83	0.50	0.67
~중간생략~												
21	0.33	0.43	0.35	0.47	0.60	0.80	0.30	0.40	0.63	0.83	0.50	0.67
22	0.83	0.66	0.83	0.66	0.70	0.56	0.85	0.68	0.75	0.60	1.00	0.80
23	0.65	0.52	0.65	0.52	0.55	0.44	0.55	0.44	0.75	0.60	0.75	0.60
24	0.78	0.78	0.78	0.78	0.65	0.65	0.75	0.75	0.63	0.63	0.75	0.75
25	0.68	0.68	0.58	0.58	0.70	0.70	0.75	0.75	0.88	0.88	0.75	0.75
26	0.70	0.56	0.65	0.52	0.50	0.40	0.75	0.60	0.75	0.60	0.75	0.60
27	0.70	0.56	0.70	0.56	0.45	0.36	0.90	0.72	0.75	0.60	0.75	0.60
28	0.33	0.33	0.33	0.33	0.55	0.55	0.30	0.30	0.88	0.88	0.50	0.50
29	0.45	0.45	0.45	0.45	0.60	0.60	0.45	0.45	0.75	0.75	0.50	0.50
30	0.48	0.63	0.48	0.63	0.85	1.13	0.65	0.87	0.75	1.00	0.75	1.00
평균	**0.56**	**0.57**	**0.58**	**0.58**	**0.61**	**0.64**	**0.62**	**0.63**	**0.66**	**0.68**	**0.71**	**0.72**

현율이 <표 1>에 요약되어 있다. 또한, 정확도와 재현율을 통합하여 측정한 F-measure의 키워드 집합 개수에 따른 성능 변화 양상은 <그림 9>를 통해 살펴볼 수 있다.



<그림 9> 키워드 집합 개수 증가에 따른 F-measure 값 향상

평가 결과, 키워드 집합의 개수는 IVSM의 성능에 상당한 영향을 미침을 알 수 있으며, IVSM의 성능을 보다 향상시키기 위해서 키워드 집합에 가능한 모든 키워드들을 포함시킬 수 있도록 유지 관리되는 것이 중요하다는 점을 유추할 수 있다.

4.2.2 입력 문서의 크기에 따른 성능 평가

입력 문서의 크기에 대한 영향을 평가하기 위해서, 유통 및 물류의 키워드 집합 830개를 축적한 뒤, 입력 문서의 크기를 다양화하여 실험을 수행하였다. 즉, 한국항만경제학회지의 최근 논문 15개에 대해서 초록, 초록과 서론, 전문(full pa-

per)의 세 가지로 크기를 다양화한 입력 문서를 만들었다.[3] 초록은 7줄 내외, 초록과 서론은 2장 이내, 그리고 전문은 20장 내외의 분량이었으며, 각각 입력 문서에 대해 생성된 키워드의 정확도를 비교했다. 그 결과가 <표 2>에 정리되어 있다.

<표 2> 입력 문서 크기 변화에 따른 정확도 비교

D#	초록	초록+서론	전문
1	0.75	0.75	0.50
2	1.00	1.00	0.75
3	0.50	0.50	0.50
4	0.50	0.75	0.75
5	0.75	0.50	0.50
6	0.75	0.75	0.75
7	0.75	0.50	0.75
8	0.75	0.75	0.75
9	0.50	0.75	0.75
10	0.75	0.50	0.75
11	1.00	1.00	0.50
12	0.75	0.50	0.75
13	0.75	0.75	0.50
14	0.50	0.50	0.25
15	0.50	0.50	0.50
평균	**0.70**	**0.67**	**0.62**

3) 키워드 생성을 위한 입력 문서는 파싱 가능한 텍스트 형식의 문서여야 한다. 실험에 사용된 한국항만경제학회지의 논문들은 원문이 '텍스트 pdf' 형식의 파일로, 원본 pdf 파일이 텍스트 파일로 쉽게 변환 가능하다. 반면, 다른 학회지의 논문들은 스캐닝을 통해 만들어진 '이미지 pdf' 형식의 파일로 텍스트 파일로 만들기 위해서는 수작업으로 입력하는 과정이 필요하다. 이에 본 실험에서는 실험의 편의 상 한국항만경제학회지의 논문들만을 사용하여 입력 문서의 크기를 다양화하는 실험을 진행했다. 이외의 실험에서는 한국유통경영학회지, 한국로지스틱스학회지, 한국물류학회지, 한국항만경제학회지, 한국해운물류학회지의 논문들 초록을 모두 입력 문서로 사용했으며, 이때 '이미지 pdf' 형식의 논문들의 경우 수작업으로 입력하여 입력 문서를 만들었다.

실험 결과에 따르면, 입력되는 문서의 크기는 성능에 큰 영향을 미치지 못함을 보여준다. 오히려 전문보다는 초록을 입력하여 키워드를 생성하는 경우 더 나은 결과를 보인다고 할 수 있다. 입력되는 문서의 크기보다는 입력되는 문서가 포함하고 있는 용어나 내용이 얼마나 응집력 있는가가 IVSM의 성능에 보다 큰 영향을 미침을 알 수 있다.

4.2.3 다양한 분야의 적용 가능성

IVSM은 '웹 기반 커뮤니티 서비스의 키워드 생성 시스템'과 '독립시스템'의 두 가지 시스템으로 구현되었다. 독립시스템에 유통 및 물류 분야의 키워드를 생성하도록 시스템이 완성되었다. 이것은 신변잡기적인 글이 아닌 특정 분야로 한정 지을 수 있고 전문성을 가진 분야의 시스템이라고 할 수 있다. IVSM이 신변잡기적인 문서와 전문성이 있는 분야 모두에 적용가능한지를 평가하기 위해, 앞 절에 소개된 유통 및 물류 분야의 시스템 외에, 다양한 유저들이 다양한 관심사를 가지고 모여 활동하는 웹 기반 커뮤니티 서비스의 신변잡기적인 글에 대한 키워드 생성 시스템이 추가적으로 구현되었다.

신변잡기적인 글들에 대한 IVSM의 성능 평가를 위해 커뮤니티 서비스에 입력된 글들 30개를 이용하여 정확도를 평가했다. 이 글들은 일반 사용자에 의해 키워드가 수작업으로 입력이 되어 있으며 입력된 키워드 역시 글의 내용과 관련되었다고 판단되는 적절한 키워드가 3~5개 지정된 것들로 선별됐다. 선별된 글들은 신변잡기적인 성격을 위해 영화, 요리, 건강 등과 관련된 것들이다. 그리고 앞서 평가된 유통물류분야 논문 키워드 생성 시스템에 대해서 30개의 초록을 가지고 그 성능을 비교했다. IVSM을 통해 생성된 키워드의 정확도가 우수한지를 객관적으로 평가하기 위해 기본시스템(baseline system) 과 비교 평가를 수행했다.

<표 3> 다양한 주제의 커뮤니티 서비스와 유통물류의 IVSM의 정확도 비교

D#	커뮤니티 서비스 baseline	커뮤니티 서비스 IVSM	유통물류 baseline	유통물류 IVSM
1	0.25	0.75	0.00	0.75
2	0.50	1.00	0.50	1.00
3	0.50	0.50	0.25	0.75
4	0.50	0.75	0.00	0.50
5	0.50	1.00	0.25	0.75
6	0.50	1.00	0.00	0.75
7	0.25	0.50	0.25	0.50
8	0.50	1.00	0.25	1.00
~ (중간생략)				
27	0.25	0.75	0.75	0.75
28	0.25	0.50	0.25	0.50
29	0.50	0.50	0.50	0.50
30	0.25	0.50	0.25	0.75
평균	**0.46**	**0.75**	**0.37**	**0.71**

기본시스템은 확률 기반의 용어빈도(term frequency)에 기초하여 키워드를 생성해내는 시스템으로 문서를 파싱하여 불용어 등을 제외한 용어들 중에 문서 출현 가능성이 높은 상위 빈도수를 갖는 용어를 키워드로 생성해내는 시스템이다. 실험결과는 <표 3>에 보이듯이, IVSM은 전문적이거나 신변잡기적인 분야 어디에나 70% 이상의 정확도를 보이며, 기본시스템과 비교한 결과 역시 정확도가 상당히 향상되었다.

4.2.4 영어 문서를 바탕으로 한 기존 시스템과의 성능 비교 평가

IVSM은 키워드 자동 할당을 위해 개발 되었으며, 특히 국문 문서의 키워드 생성에 적합하도록 최적화되었다. 아직까지 국문 문서의 키워드 자동 생성을 위한 툴들이 개발되지 않았으므로 IVSM의 성능과 비교를 위한 모델이나 시스템이 전무하다. 이에 본 연구에서는 해외에서 개발된 기계학습 기법을 적용한 대표적인 키워드 자동 추출 툴인 GenEx의 공개 버전인 Extractor와 비교 평가를 수행했다. 현재 Extractor 7.2[4] 버전이 공개되었다.

Extractor와의 비교평가는 영문 문서를 바탕으로 수행되어야 한다. 이를 위해 본 실험에서는 Data and Knowledge Engineering 저널 Vol. 30부터 Vol. 60까지의 논문들에서 총 435개의 키워드 집합을 수집했다. 키워드 자동 할당을 위한 입력 문서는 Vol. 29와 Vol. 61에 포함된 논문 중 10편의 초록과 서론을 평가데이터로 사용했다. 평가 데이터 논문에 저자가 할당한 키워드 개수는 평균적으로 5개였으므로, 비교를 위해 Extractor와 IVSM 각각에서 생성된 키워드의 개수 역시 5개로 제한하였다. 실험 결과는 다음의 <표 4>에 정리되어 있다.

<표 4> Extractor vs. IVSM 성능 비교

D#	Extractor		IVSM	
	정확도	재현율	정확도	재현율
1	0.33	0.40	0.33	0.40
2	0.67	0.67	0.67	0.67
3	0.50	0.60	0.50	0.60
4	0.67	0.80	0.50	0.60
5	0.50	0.50	0.33	0.33
6	0.33	0.50	0.33	0.50
7	0.50	0.75	0.50	0.75
8	0.50	0.75	0.50	0.75
9	0.50	0.60	0.50	0.60
10	0.67	0.57	0.83	0.71
평균	**0.52**	**0.61**	**0.50**	**0.59**

4) http://www.extractor.com.

실험 결과, Extractor가 IVSM보다 다소 좋은 성능을 보이는 것으로 평가되었다. 또한 현재 IVSM은 국문 문서 키워드 생성에 최적화되어 있으므로 영문 문서에 적용한 결과는 국문 문서에 적용했을 때 보다 낮은 성능을 보이고 있다.

4.3 평가 결과 논의

IVSM의 성능 평가는 (1)축적된 키워드 집합의 개수에 대한 의존도, (2)입력 문서의 크기, (3)다양한 분야의 적용 가능성, 그리고 (4)영문 문서를 바탕으로 한 기존 시스템(Extractor)과의 성능 비교의 네 가지 측면에 대해 수행되었다.

첫 번째로, 키워드 집합의 개수에 대한 의존도에 대해서 평가해 본 결과 IVSM은 키워드 집합의 수가 증가할수록 정확도가 높아짐을 알 수 있다. 키워드 생성의 두 가지 방법 중 본 연구에서는 키워드 할당 방법을 취하고 있는데, 실험 결과로 비추어 이 키워드 할당 방법에서 중요시 다루고 있는 미리 정의된 사전 혹은 어휘집의 중요성을 다시 한번 확인할 수 있었다. 키워드 할당 방법은 키워드 생성을 위해 사용되는 사전을 얼마나 잘 정의해 놓느냐가 핵심이 될 수 있다. IVSM 역시 사전에 해당되는 키워드 집합을 얼마나 많이 그리고 양질의 키워드 집합을 축적해 놓느냐가 성능에 큰 영향을 미쳤다. 향후 IVSM을 적용하여 지속적으로 키워드 집합을 생성하고 이를 축적하면 보다 많은 키워드 집합이 쌓이게 될 것이며, 이는 결론적으로 IVSM의 성능을 점차적으로 향상시킬 수 있는 가능성이 있음을 보여준다.

두 번째로, 입력 문서의 크기에 대한 IVSM의 성능평가를 위해, 입력 문서를 초록, 초록과 서론, 그리고 전문(full paper)으로 입력 문서의 크기를 변화시켜 실험을 수행했다. 그 결과 문서의 사이즈는 정확도에 큰 영향을 미치지 않는 것으로 분석되었다. 오히려 전문을 사용한 결과 관련이 없는 키워드를 생성해낼 가능성이 높아지는데, 이는 전문에 포함된 용어가 특정 주제나 용어로 응집되지 못하는 경우가 많음에 기인한 것으로 추측할 수 있다. 반면, 논문 초록의 경우 비록 많은 용어를 포함하고 있지는 않지만 논문의 내용을 압축해서 요약하고 있으며, 논문의 내용과 관련된 비중 있는 용어들이 선별되어 집합되어 있기 때문에 좋은 결과를 보여주는 것으로 생각된다. 그러나, 만약 문서의 분량이 길어진다면, *tf*(term frequency)에 기초한 가중치 부여 보다는 *tfidf*(term frequency inverse document frequency)와 같은 상대적 빈도수를 기초로 한 가중치 부여 방식을 적용하여 덜 중요한 빈발용어에 대한 처리가 가능할 것이다.

세 번째로 다양한 분야의 적용 가능성에 대한 평가를 위해 신변잡기적인 글들에 대한 시스템인 커뮤니티 서비스 IVSM과 유통물류분야를 위한 독립 IVSM의 성능평가 결과 각각 75%와 71%의 정확도를 보였다. 이는 IVSM이 비교적 짧고, 신변잡기적인 성향의 글들이 있는 웹 기반 커뮤니티 서비스나, 물류와 같은 전문성이 있는 분야 모두 기본시스템(baseline system)보다 좋은 성능을 보였다. 다시 말해서, IVSM은 특정 분야의 전문적인 정도와는 무관하게 다양한 분야에 적용 가능함을 시사한다.

마지막으로, 영문 문서를 사용하여 Extractor와의 성능 평가 결과 Extractor가 IVSM보다 다소 좋은 성능을 보이는 것으로 평가되었다. 또한 IVSM은 국문 문서보다 영문 문서의 성능이 떨어졌다. 이는 당연한 결과로, 파싱 알고리즘 및 복합어 처리 등에 있어 IVSM이 국문 문서에 적합하도록 최적화되어 있기 때문이다. 그럼에도 불구하고 IVSM이 Extractor와도 견줄만한 좋은 결과를 보이고 있다는 것은 주지해야 할 점이다. 즉, IVSM이 키워드 자동 생성을 위한 모델로서 특정 언어에 의존적이지 않고 다양한 언어에 적용될 수 있음을 보여준다.

V. 결론 및 향후 연구과제

최근 디지털 도서관이 등장하고 인터넷이 폭넓게 보급되어 온라인 상에서 얻을 수 있는 텍스트 정보의 양이 급증하면서 문서 관리의 효율과 검색 성능의 향상을 꾀하기 위해 키워드 할당의 필요성에 대해 의견을 같이 하고 있다. 사람은 문서를 읽고 그 내용을 머릿속에서 개념적으로 정리하여 몇 개의 키워드들을 인지한다. 그러나 이와 같은 과정을 사람이 스스로 사고하여 완성한다는 것은 시간적, 비용적으로 비효율적이다. 따라서 이 과정을 개선하기 위하여 문서를 대표할 수 있는 주요 키워드를 자동으로 생성하는 방법에 관해 연구들이 진행되어 왔다. 특히, 인간이 학습하는 과정을 모방하여 기계적 시스템으로 하여금 학습을 통해 추론, 연산, 판단하도록 하는 기계학습 분야에서 키워드 자동 생성에 관심을 갖고 몇몇 연구가 실험적으로 진행되기도 했다.

그러나 아직까지는 키워드 자동 생성 자체의 성능을 향상시키기 위해 키워드 생성 시스템에 초점을 맞춰 연구를 진행한 것은 일부 연구에 지나지 않는다. 대부분이 문서 범주화나 문서 요약을 위한 부분연구로서 키워드 자동 생성에 대한 연구를 접근했다. 그러나 문서 범주화나 문서 요약을 위해 키워드를 생성하는 것은 궁극적으로 그것을 위한 시스템의 성능 향상을 위해 디자인되기 때문에, 실제 문서 내용에 근거하여 논문이나 도서의 키워드를 생성하는데 적용할만한 키워드 자동 생성기의 성능에는 부합되지 못하는 경우가 많다. 따라서 다른 시스템을 위한 서브모듈로서 키워드 생성기가 아닌 본연이 목적 자체가 키워드 자동 생성을 위한 방법론 및 시스템에 대한 연구에 보다 관심을 가질 필요가 있다.

이에 본 연구는 키워드 자동 생성을 위한 IVSM을 제안했다. IVSM은 입력된 문서에 대해서 가장 근접한 키워드 집합을 찾아내어 이를 바탕으로 키워드를 생성하는 새로운 방법으로서, IVSM은 글을 입력하거나 문서를 공유하는 다양한 웹 기반 서비스 혹은 전자적인 형태로 존재하는 모든 문서 등의 키워드 생성에 직접적으로 적용 가능하다. 실제 평가결과 웹 기반 커뮤니티 서비스 IVSM과 물류분야를 위한 독립 IVSM의 성능평가 결과 각각 75%와 71%의 정확도를 보였다. 이는 비교 평가를 위해 구현된 기본 시스템(baseline system) 보다 거의 2배 이상 높은 정확도이다. 특히, 영문 문서에 최적화되지 않은 IVSM이 영문 문서를 위한 키워드 자동 추출 툴인 Extractor에 견줄만한 성능을 보인다는 것은 IVSM이 최적화 과정을 통해 영어를 포함한 다양한 외국어 문서에 적용했을 또한 좋은 결과를 보여줄 것으로 기대할 수 있다. 또한, 이 정확도들은 다음의 향후 연구를 통해 보완되어 향상될 수 있을 것으로 기대된다.

첫째, 키워드 집합의 정제 및 다수의 키워드 집합 축적을 통한 키워드 사전의 완성도를 높임으로써 정확도를 보다 향상 시킬 수 있다. 본 연구에서는 일부 불용어의 제거 등 최소한의 정제 과정만을 거친 키워드 집합을 사용했으며, 유통물류 분야 키워드 역시 국내 관련저널의 가능한 거의 모든 키워드 집합을 축적하긴 했으나, 아직 수적으로 부족한 것이 사실이며, 향후, 키워드 집합의 수를 지속적으로 증가시켜 키워드 생성 시스템의 성능을 높일 수 있을 것이다. 또한 충분한 키워드 집합을 가지고 있다면 얼마만큼의 키워드 집합의 개수에서 수렴하는 지에 대한 결과도 확인할 수 있을 것이다.

둘째, 키워드 생성을 위해서는 입력 문서가 파싱되어 문서를 구성하고 있는 용어들에 대한 형태소 분석이 이루어져야 한다. 한글 형태소 분석 시스템 개발을 위한 다양한 노력이 있어왔음에도 불구하고 아직까지는 그 성능에 한계가 있다. 특히, 본 연구에서는 루씬 한글 형태소 분석 알고리즘을 적용했으나, 이것은 영어에 맞게 만들어진 분석 알고리즘을 한글에 맞춰 개선한 것이기 때문에 만족할 만한 결과를 보여주지 못하고 있다. 영어보다 우리말은 다양한 조사와 복합어로

인해서 형태소 분석이 용이하지 않은 것이 사실이며, 그렇기 때문에 한글 형태소 분석의 만족할 만한 결과를 보여줄 수 있다면, 본 연구에서 제안한 모델의 성능은 보다 개선될 여지가 충분히 있다.

마지막으로, 본 연구에서는 전통적인 벡터공간 모델에 사용되는 용어 가중치를 그대로 적용하였으나, 여러 가지 용어 가중치를 변형 적용하여 IVSM에 적합한 용어 가중치 방법을 선정 혹은 개발함으로써 IVSM의 성능을 개선할 수 있을 것이다.

〈References〉

[1] Berry, M., Dramc, Z., and Jessup, E., "Matrics, Vector Spaces, and Information Retrieval," *SIAM Review*, Vol. 41, 1999, pp. 335-362.

[2] Ercan, G. and Cicekli, I., "Using lexical chain for keyword extraction," *Information Processing and Management*, Vol. 43, 2007, pp. 1705-1714.

[3] Frank, E., Paynter, W., Witten, I., Gutwin, C., and Nevill-Manning, C., "Domain-specific keyphrase extraction," In: *Proceedings of IJCAI'99*, 1999.

[4] Kim, P., "A Study on the Performance Improvement of Rocchio Classifier with Term Weighting Methods," *Journal of the Korean Society for information Management*, Vol. 25, No. 1, 2008, pp. 211-233.

[5] Lee, C., Kim, M., Lee, K., Lee, G., and Park, H., "Document Thematic words Extraction using Principal Component Analysis," *Journal of Korean Institute of Information Scientists and Engineers*, Vol. 29, No. 10, 2002, pp. 747-754.

[6] Lee, M. and Bae, H., "Design of Keyword Extraction System Using TFIDF," *Korean journal of cognitive science*, Vol. 12, No. 1, 2002, pp. 1-11.

[7] Matsuo, Y. and Ishizuka, M., "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, Vol. 13, No. 1, 2004, pp. 157-169.

[8] Salton, G. and McGrill, M., *Introduction to modern information retrieval*, Mcgraw-Hill, New York, 1983.

[9] Shin, H., Zhang, B., and Kim, Y., "Learning Probabilistic Graph Models for Extracting Topic Words in a Collection of Text Documents," In: *Proceedings of Spring Conference on Korean Institute of Information Scientists and Engineers*, 2000.

[10] Turney, P., "Extraction of Keyphrase from text: evaluation of four algorithms," *National Research Council, Institute for Information Technology*, Technical Report ERB-1051, 1997.

[11] Turney, P., "Learning to Extract Keyphrases from Text," *National Research Council, Institute for Information Technology*, Technical Report ERB-1057, 1999.

[12] Turney, P., "Learning algorithm for keyphrase extraction," *Information Retrieval*, Vol. 2, No. 3, 2000, pp. 303-336.

[13] Wan, X., "A novel document similarity measure based on earh mover's distance," *In_formation Sciences*, Vol. 177, 2007, pp. 3718-3730.

[14] Witten, I., Paynter, G., Frank, E., Gutwin, C., and Nevill-Manning, C., "KEA: Practical Automatic Keyphrase Extraction," In: *Proceedings of DL'99*, 1999.

[15] Yang, K. and Huh, S., "Automation of Expert Classification in Knowledge Management Systems Using Text Categorization Techni-

que," *Journal of MIS research*, Vol. 14, No. 2, 2004, pp. 115-130.

[16] Yoon, A., Hwang, S., Lee, E., and Kwon, H., "Construction of Korean Wordnet KorLex 1.5," *Journal of Korean Institute of Information Scientists and Engineers*, Vol. 36, No. 1, 2009, pp. 92-108.

[17] Zhang, C. and Xu, H., "Using citation-KNN for automatic keywords assignment," In: *proceedings of 2009 International Conference on Electronic Commerce and Business Intelligence*, 2009.

◆ About the Authors ◆

Wonchin Cho

Wonchin Cho received the B.S. and M.S. degree in Business from Ajou University in 2001 and 2003. She received her Ph.D. from Seoul National University in 2010. Currently, she is a researcher of Institute of Management Research at Seoul National University, and an adjunct Professor of E-Business at Ajou University. Her research interests include data mining, Semantic Web, ontology, and automatic keyword generation.

Sangkyu Rho

Sangkyu Rho is Professor of Information Systems in the Graduate Schoolof Business at Seoul National University. He received his Ph.D. from the University of Minnesota. His research interests include Internet business, ontology development, data mining, ranking, and social networking. He has published papers in such journals as IEEE Transactions on Knowledge and Data Engineering, Information Systems, Annals of Operations Research, and Strategic Management Journal.

Jiyoung Agnès Yun

Jiyoung Agnès Yun works for NCSoft as director of new services. Yun earned her PhD in Sociology of Communication from Paris 5 University (Université de René Descartes, Sorbonne) with a dissertation on *Intervention of new network in the social identity construction of European Union.* After coming back to Korea, she then founded Media IN Labin 2002 where she conducted various strategic consulting projects in the areas of the Internet, public space, telecommunication, and social interaction. Yun then served as Director of R&D Institute and vice president of SK Communications where she developed new service platforms for wireless and wired Internet. She then founded an Internet Start-up Mediare Inc., where she launched *Itgling*, a social networking service based on hybrid-link.

Jinsoo Park

Jinsoo Park is Associate Professor of Information Systems in the Graduate School of Business at Seoul National University. He was formerly on the faculties of University of Minnesota and Korea University. He holds a Ph.D. in MIS from the University of Arizona. His research interests include ontology, semantic interoperability, metadata management, data modeling, and process modeling. His research has been published in MIS Quarterly, IEEE Transactions on Knowledge and Data Engineering, IEEE Computer, ACM Transactions on Information Systems, and others. He currently serves on the editorial boards of Journal of Database Management and International Journal of Principles and Applications in Information Science and Technology.