정보관리연구, vol.38, no.1, 2007, pp.1-19

데이터 결합이 웹 문서 검색성능에 미치는 영향 연구

A Study on the Effect of Data Fusion on the Retrieval Effectiveness of Web Documents

박 옥 화*·정 영 미**
Ok-Hwa Park · Young-Mee Chung

차 례

초 록

이 연구에서는 최근 검색성능을 향상시키기 위한 전략으로 사용되는 데이터 결합기법을 웹 문서 검색에 적용하고, 실험을 통해 문서표현 방법의 결합이 검색성능에 미치는 영향을 분석하였다. 문서표현 방법으로는 내용기반 표현, 링크기반 표현, URL 등을 선정하고, 단일 표현 방법에 의한 검색결과와 표현방법의 결합을 통한 검색결과를 비교하였다. 분석결과 다른 문서표현 방법의 결합이 웹 문서의 검색성능을 향상시키지는 못하는 것으로 나타났다.

키 워 드

정보검색, 데이터 결합, 웹 문서, 문서표현, 검색성능

* 다음 커뮤니케이션 검색포털본부 통합검색팀 팀원
(Search Portal Division, Search Service Team Member, Daum Communications, parkokhwa@hanmail.net)
** 연세대학교 문헌정보학과 교수
(Professor, Dept. of Library and Information Science, Yonsei University, ymchung@yonsei.ac.kr)
• 논문접수일자 : 2006년 11월 17일
• 게재확정일자 : 2007년 3월 8일

ABASTRACT

This study investigates the effect of data fusion on the retrieval effectiveness by performing an experiment combining multiple representations of Web documents. The types of document representation combined in the study include content terms, links, anchor text, and URL. The experimental results showed that the data fusion technique combining document representation methods in Web environment did not bring any significant improvement in retrieval effectiveness.

KEYWORDS

Data Fusion, Web Documents, Information Retrieval, Document Representation, Retrieval Effectiveness

1. 서 론

정보검색에 대한 연구는 1950년대 미국에서 본격적으로 시작되었으며, 1990년대 들어 인터넷과 웹의 확산으로 인해 이에 대한 관심이 더욱 커지고 있다. 특히 웹 검색엔진을 통한 정보검색의 대중화는 일반 이용자의 정보요구를 만족시킬 수 있는 지능적 시스템에 대한 필요성과 검색성능의 향상을 위한 새로운 기법에 대한 요구를 증대시키고 있다.

검색성능을 향상시키기 위한 전략의 하나인 데이터 결합(data fusion)은 여러 다른 검색모형을 결합하거나 다른 유형의 색인어들을 결합함으로써 검색성능을 높이려는 전략이며, 단일 검색모형이나 단일 문서표현 기법을 사용하여 얻을 수 있는 검색성능의 한계를 극복하고자 하는 시도이다(정영미 2005).

데이터 결합은 다음과 같은 가설에 기반을 두고 있다. 즉 상이한 검색관련 기법에 의해 검색된 문서들은 상호 중복성(overlap)이 낮으며(Belkin et al. 1993), 적합문서들은 여러 다른 검색관련 기법들에 의해 공통적으로 검색되지만 부적합문서들의 검색결과는 기법에 따라 달라지는 경향이 있다는 것이다(Lee 1997; Vogt and Cottrell 1998). 따라서 다양한 여러 기법을 사용하여 검색한 결과를 통합하면 단일 기법을 사용하는 것보다 질의에 적합한 문서를 더 많이 제공할 수 있을 것이라고 보는 것이다.

지금까지 데이터 결합에 관한 연구는 다른 유형의 색인어의 결합, 색인어 추출 대상 텍스트의 결합, 용어 가중치의 결합, 질의형식의 결합, 검색모형의 결합, 검색 순위의 결합 등 다양한 측면에서 수행되어 왔으며, 대체적

으로 데이터 결합 결과 성능이 향상된 것으로 나타났다(McGill and Noreault 1979; Katzer et al. 1983; Turtle and Croft 1991; Fox and Shaw 1993, 1994; Salton, Allan, and Buckley 1993; Callan 1994; Belkin et al. 1995; Lee 1995; Rajashekar and Croft 1995; Vogt and Cottrell 1999; Croft 2000; 최성환 2001; 전상우 2005).

최근에 들어서는 웹 문서를 대상으로 하여 다양한 기법들을 결합하는 연구가 수행되고 있다(Westerveld, Kraaij, and Hiemstra 2001; Yang 2001; Tsikrika and Lalmas 2002, 2004; 안동언, 강인호 2003; Ogilvie and Callan 2003; Kang and Kim 2003). 이 연구들에서는 주로 내용기반 표현과 링크기반 표현의 결합을 통해 검색성능의 향상을 시도하였으나, 두 가지 표현의 결합이 내용 정보만을 사용했을 경우에 비해 눈에 띄는 성능향상을 가져오지는 않았다.

정보검색 콘퍼런스인 TREC(http://trec.nist.gov/)의 웹 트랙에서는 지정 페이지/사이트 검색, 토픽 검색, ad-hoc 검색 등 세 가지 검색 과제에 대해 많은 실험들이 수행되었는데, 실제로 지정 페이지/사이트 검색 과제에서는 이러한 결합이 성능의 향상을 가져온 반면, ad-hoc 검색에 있어서는 검색성능에 거의 영향을 미치지 못하는 것으로 나타났다(Hawking et al. 1999; Hawking 2000; Hawking and Craswell 2001).

뿐만 아니라 TREC 실험집단에 따라서도 그 결과가 상이하게 나타나고 있는데 대표적인 실험집단인 WT2g와 WT10g를 가지고 동일한 실험을 수행하여도 같은 결과가 나타나지 않고 있다. 또한 질의의 특성에 따라서도 결과가 달라지는 것으로 나타났는데 동일한 실험집단과 동일한 주제문을 가지고 수행하는 실험의 경우라도 질의에 사용되는 가중치, 생성된 질의의 길이 등에 따라서 상이한 결과를 가져오는 것을 볼 수 있다.

이와 같이 웹 문서 검색에서 데이터 결합을 통해 검색효율을 향상시키고자 하는 노력이 있어 왔지만 이러한 결합이 검색과제, 실험집단, 질의특성 등에 따라 상이한 결과를 나타내고 있기 때문에 이론적으로 데이터 결합의 효율성을 명료하게 입증하기는 어려운 것으로 보인다.

이 연구에서는 추론망 검색모형의 결합 가중치에 의해 웹 문서 표현기법의 결합수준을 달리함으로써 최적의 데이터 결합이 검색성능의 향상을 가져올 수 있는 지를 알아보고자 한다. 이를 위해 문서표현을 내용, 링크, URL의 세 영역으로 나누고, 이러한 웹 문서 표현방법의 결합이 검색성능에 어떠한 영향을 미치는 지를 평가한다.

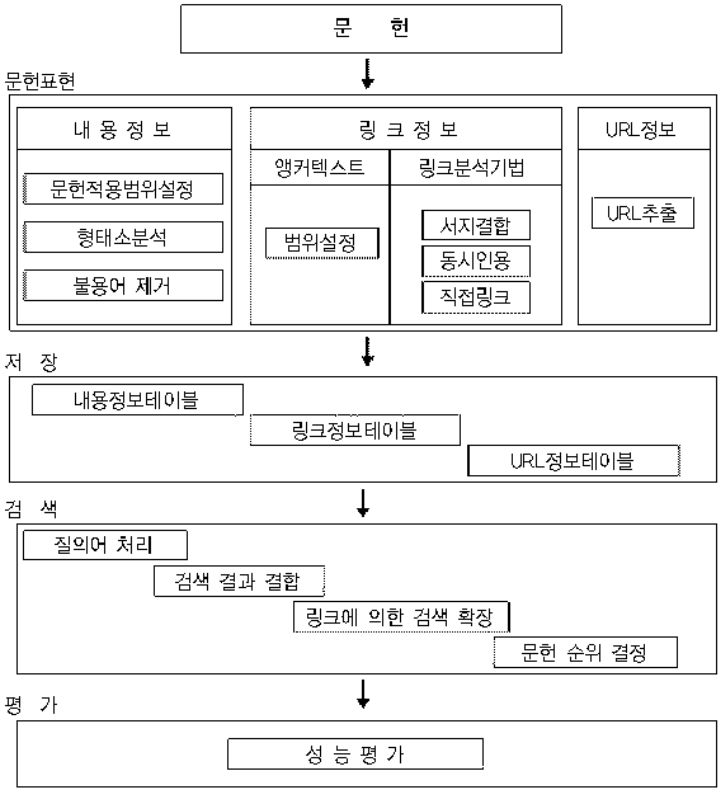
검색실험을 위해 펜티엄 IV에 윈도우 운영체제를 기반으로 하는 실험시스템을 구축하고 데이터베이스로는 MySQL을 이용하였다. 실험집단으로는 TREC에서 웹 실험집단으로 제공하는 WT2g를 사용하였으며, 검색모형으로는 문서의 다양한 표현방법의 결합을 지원하는 추론망 모형을 채택하였다.

2. 문서표현 결합 실험 설계

2.1 실험 개요

이 연구에서는 웹 문서의 검색성능을 높이기 위한 방법으로 내용 정보와 링크 정보, URL 정보를 결합하는 검색실험을 수행하였으며, 검색모형으로는 추론망 모형을 사용하였다(Turtle 1990). 〈그림 1〉은 이 연구에서 수행된 실험의 개요를 보여 준다.

문서의 내용기반 표현은 〈title〉 태그만을 이용한 표현과 〈title〉 태그에 〈font〉, 〈h〉 태그가 추가된 표현으로 구분하였다. 이전에 수행되었던 연구들은 〈title〉 태그만을 이용하거나(Trikrika 2002, 2004), 〈body〉 태그 전체를 이용하였는데(Yang 2001; Kraaij et al. 2002), 〈title〉 태그만을 이용할 경우에는 웹 문서의 내용을 표현하기에 부족할 뿐만 아니라 웹 문서의 저자가 〈title〉 태그에 크게 비중을 두지 않음으로써 특정 홈페이지의 모든 웹 문서에 동일한 〈title〉을 부여한 경우가 존재한다. 반면 〈body〉 태그 전체를 이용할 경우 웹 문서의 내용을 좀 더 정확하게 표현할 수는 있지만, 매우 많은 양의 웹 문서로 인해



〈그림 1〉 실험 개요

데이터 처리에 어려움이 있다.

〈font〉 태그와 〈h〉 태그를 사용한 이유는 웹 문서 작성 시 글씨의 크기나 색상을 바꾸어 중요한 내용을 표현하거나 제목에 주제어를 사용하는 경향이 있기 때문이다. 본 연구에서는 먼저 〈title〉 태그와 함께 〈font〉 태그와 〈h〉 태그의 효용성을 평가한 다음 성능이 좋은 내용기반 표현방법을 다른 표현방법과 결합하였다.

링크기반 표현에는 앵커텍스트(anchor text), 직접링크(direct link), 동시링크(co-link)의 세 가지 표현방법이 사용된다. 동시링크 기법은 문서 인용에서의 서지결합 기법 및 동시인용 기법에 대응되는 개념으로 동시-아웃링크(co-outlink)와 동시-인링크(co-inlink)가 이용된다(정영미 2005; Thelwall and Wilkinson 2004; Tsikrika and Lalmas 2004).

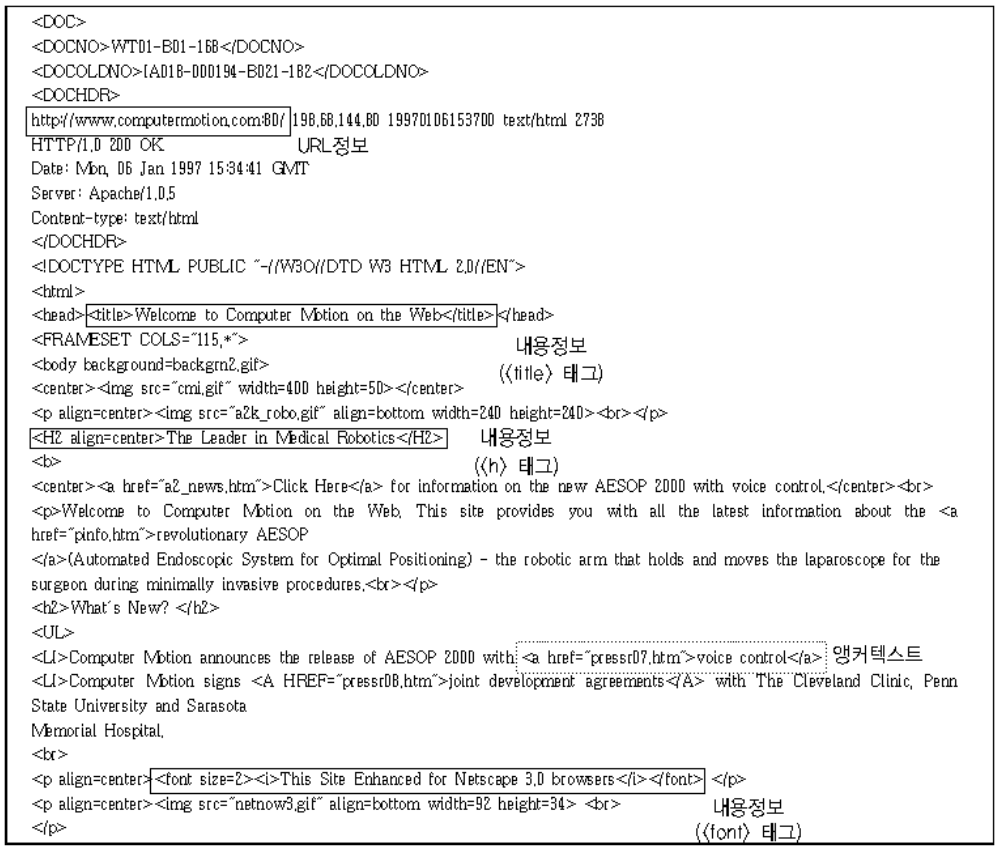
앵커텍스트는 전체 웹 문서 텍스트에 포함된 용어들보다 상대적으로 중요한 용어를 사용함으로써 웹 문서와 실질적으로 관련 있는 다른 웹 문서에 대한 정보를 제공하는 것으로 알려져 있다(Bharat and Henzinger 1998; Amitay 1998; Chakrabarti et al. 1998; Davison 2000). 이 연구에서는 앵커텍스트에 의해 링크를 받은 웹 문서가 실험문서집단에 거의 포함되어 있지 않았기 때문에 앵커텍스트를 링크가 출현한 웹 문서의 자질로 사용하였다.

URL 정보는 웹 문서를 작성한 기관의 이름이나 약어를 포함할 수 있기 때문에 홈페이지의 첫 화면을 찾는 검색작업에서 주로 적용되어 왔다(Kang and Kim 2003; 안동언, 강인호 2002). URL은 질의어와 직접 비교하는 방식과 URL 형태에 따라 사전확률(URLprior)을 구하여 이를 이용하는 방식으로 구분될 수 있는데 본 실험에서는 URL을 질의어와 비교하는 방식을 택하였다.

〈표 1〉은 실험에서 사용한 내용, 앵커텍스트, 링크, URL 등 네 가지 문서표현 방법에 대해 정리한 것이고, 〈그림 2〉는 실험대상인

〈표 1〉 실험에 사용된 문서표현 방법

문헌표현	구 분	상세내용	구분기준
내용		〈title〉 태그	텍스트 적용 범위
		〈title〉, 〈h〉, 〈font〉	
앵커텍스트		〈a〉〈/a〉사이에 있는 앵커텍스트	
링크	직접링크	웹 문서의 링크를 직접 이용	링크 이용 방식
	동시인용	인링크를 이용	
	서지결합	아웃링크를 이용	
URL			



〈그림 2〉 실험문서의 예

웹 문서에서 직접적으로 추출되는 내용정보를 표시한 예를 보여 준다.

2.2 실험문서집단

웹 문서의 다양한 문서표현 방법이 정보검색 성능에 미치는 영향을 검증하기 위한 실험문서집단은 TREC의 웹 문서 실험집단인 WT2g로부터 구축하였다. TREC-8부터 사용되기 시작한 WT2g는 일반적인 검색과제(ad-hoc retrieval)를 위한 실험집단으로 많은 검색실험에서 사용되어 왔기 때문에 신뢰성이 보장되어 있다.

WT2g는 총 956개의 호스트로 구성되어 있으며, 문서 수는 24만7,491개, 평균 적합문서 수는 45.58개이다. 그리고 실험문서와 함께 제공되는 질의는 총 50개로 짧은 질의어 대신 주제문(topic statement) 형태로 제공하며, 이로부터 연구자들의 임의로 질의를 구성할 수 있도록 하고 있다. 주제문은 〈title〉, 〈description〉, 〈narrative〉 필드로 구성되어 있는데 〈title〉 필드는 질의를 잘 표현하는 핵

심 단어이고, 〈description〉 필드는 〈title〉 필드에 대한 좀 더 구체적인 설명을 포함하고 있으며, 〈narrative〉 필드는 적합문서 판단의 기준을 나타내고 있다.

본 연구에서는 WT2g의 일부 문서만을 대상으로 하였는데, 그 이유는 약 25만개의 문서를 처리하기에는 시스템의 한계가 존재하기 때문이었다. 따라서 임의로 선택한 질의 20개에 대해 내용정보의 〈title〉 태그만을 이용하여 검색된 문서들로 이 연구의 실험문서집단 WT2g-sub를 구축하였다. 이렇게 구축된 실험문서집단 WT2g-sub는 799개의 호스트, 2만695개의 문서, 평균 65.5개의 적합문서, 10개의 질의, 질의당 평균 22개의 단어 등의 통계적 특성을 갖는다.

〈표 2〉의 실험용 질의 10개는 TREC에서 제공되는 50개의 질의어 중 적합문서가 70개 이상인 질의를 선택한 것이다. 적합문서가 많은 질의어를 선택한 이유는 다양한 문서표현 방법을 결합해야 하므로 각 질의에 대한 적합문서의 수가 적거나 존재하지 않을 경우 검색결과를 평가하기 어렵기 때문이다. 각 질의는 주제문의 〈title〉 필드에 출현한 단어들로 구성하였다.

2.3 검색모형 및 가중치

추론망 검색 모형에서는 특정 문서에 출현한 각 용어에 대한 신념, $bel(t_i|D_j)$를 구한 다음, 이를 바탕으로 특정 질의에 대한 각 문서의 적합성을 나타내는 신념, $bel(Q|D_j)$를 산출함으로써 질의에 대한 문서들을 검색한다.

특정 문서에 출현한 각 용어에 대한 신념, $bel(t_i|D_j)$ 값으로는 보통 해당 용어의 TF·IDF

〈표 2〉 실험용 질의

질의번호	질 의 어	적합문서 수
402	behavioral, genetics	98
404	Ireland, peace, talks	77
407	poaching, wildlife, preserves	75
408	tropical, storms	78
412	airport, security	70
418	quilts, income	102
424	suicides	130
431	robotic, technology	148
434	Estonia, economy	110
444	supercritical, fluids	106

가중치를 이용하는데 Turtle(1990), Turtle과 Croft(1991)의 연구에서는 가중치 방식에 따라 신념 값이 달라질 수 있음을 강조하였다. 그러므로 본 연구에서는 추론망 모형을 기반으로 구현한 INQUERY 시스템의 TREC-7 실험에서 성능이 좋은 것으로 나타난 Okapi TF·IDF 가중치를 사용하여 $bel(t_i)$을 산출하였다(Allan et al. 1998). 아래의 Okapi TF·IDF 공식에서 0.4는 부모노드들이 모두 거짓일 때 $bel(t_i)$가 갖는 기본 값이다.

$$bel(t_i) = 0.4 + 0.6 \times \frac{tf}{tf + 0.5 + 1.5\frac{length}{avglength}} \times \frac{\log\frac{N}{n_i}}{\log N}$$

〈공식 1〉

tf : 특정 문서에서 출현한 단어의 빈도

$length$: 문서의 길이

$avglength$: 문서의 평균 길이

N : 전체 문서의 수

n_i : t_i가 한 번 이상 출현한 문서의 수

위의 공식을 통해 각 질의어에 대한 $bel(t_i)$가 계산되면 질의에 대한 문서의 신념 값인 $bel(Q|D_j)$를 산출한다. $bel(Q|D_j)$를 계산하는 방식은 질의 언어 연산자(query language operator)에 의해 차이가 있지만 본 연구에서는 가장 일반적인 방법인 가중치 부여 합산(weighted sum) 방법을 사용하였다.

$$bel(Q|D_j) = \frac{w_1 \cdot bel(t_1) + \cdots + w_m \cdot bel(t_m)}{w_1 + \cdots + w_m}$$

〈공식 2〉

위 공식에서 w_i는 각 질의어에 부여되는 가중치인데 이 실험에서 질의어에는 가중치를 부여하지 않았으므로 1의 값을 갖는다.

〈공식 2〉에 의해 개별 문서표현 방법에 대한 신념 값을 계산한 다음, 여러 문서 표현($\boldsymbol{R_k}$)을 결합할 때도 〈공식 3〉과 같이 가중치부여 합산방식을 이용하였다. 질의어에 가중치를 부여하지 않은 것과는 달리 문서 표현을 결합할 때는 다양한 결합 가중치를 사용하여 실험을 수행하였다.

$$bel(I|D_j) = \frac{w_1 \cdot bel(Q|R_1) + \cdots + w_k \cdot bel(Q|R_k)}{w_1 + \cdots + w_k}$$

〈공식 3〉

검색결과의 평가척도로는 일반적으로 많이 사용되는 정확률 이외에 얼마나 많은 적합문서가 상위에 검색되는 지를 평가하는 n-순위에서의 정확률을 사용하였다. 이 실험에서는 상위 5-순위, 10-순위, 20-순위에서의 정확률을 산출하여 각각 P(5), P(10), P(20)으로 표기하였다. 세 가지 순위를 선택한 이유는 웹 검색 시 이용자들은 보통 1-2페이지 정도를 훑어보며, 한 페이지는 보통 10개의 검색결과를 보여 주기 때문이다(Jansen, Spink, and Saracevic 2000; Jansen and Spink 2005).

3. 실험결과 및 평가

3.1 단일문서 표현방법을 이용한 검색 결과

먼저 단일 유형의 웹 문서 표현방법을 이용한 검색을 수행하였다. 문서 표현 방법은 내용, 앵커텍스트, URL이며, 특히 내용기반 표현의 경우 〈title〉 태그만을 이용한 검색과 〈title〉, 〈h〉, 〈font〉 태그를 동시에 이용한 검색으로 나누어 실험하였다.

〈표 3〉에는 모든 단일문서 표현방법에 대해 10개의 질의로 검색한 결과 평균 검색문서 수와 검색문서 중 평균 적합문서 수가 나와 있다. 검색결과 전반적으로 검색문서 수에 비해 적합문서의 수가 엄청나게 적었으며, 〈title〉 태그와 〈title〉, 〈h〉, 〈font〉 태그를 함께 사용한 경우의 평균 정확률이 각각 2.2% 수준에 불과하였다.

〈표 4〉는 단일 문서표현 방법으로 검색한 결과를 n-순위 정확률로 평가한 것이다.

먼저 〈title〉 태그와 〈title〉, 〈h〉, 〈font〉 태그를 함께 이용한 방법을 비교하면 평균 정확률과 평균 n-순위 정확률 모두 비슷한 값을 보이고 있으나 평균 n-순위 정확률은 세 가지 태그를 함께 사용한 경우가 다소 높았다. 세 가지 유형의 단일 문서표현 방법을 비교하여 보면, 평균 n-순위 정확률은 내용정보가 0.242와 0.253으로 가장 높고, 그 다음으로 URL 정보, 앵커텍스트의 순으로 나타났다.

내용기반 문서표현 방법은 다른 유형의 문

〈표 3〉 단일 문서표현 방법에 의한 검색결과 (평균 정확률)

단일 표현 방법	검색문서 수	적합문서 수	평균 정확률
title	406.5	9.0	0.022
title/h/font	468.8	9.3	0.020
anchor	1966.7	5.5	0.003
URL	90.6	3.0	0.033

〈표 4〉 단일 문서표현 방법에 의한 검색결과 (평균 n-순위 정확률)

	title	title/h/font	anchor	URL
P(5)	0.360	0.320	0.140	0.140
P(10)	0.220	0.250	0.100	0.150
P(20)	0.145	0.190	0.070	0.085
평 균	0.242	0.253	0.103	0.125

서표현 방법과 비교할 때 2배 이상 성능이 우수한 것으로 나타났다. 앵커텍스트는 HITS, PageRank 같은 링크 구조 분석 알고리즘과 함께 내용 정보만으로 표현한 웹 문서의 검색 성능을 높이기 위해 자주 사용되는 웹 문서의 표현방법이다. 하지만 앵커텍스트를 이용한 문서표현 방법의 경우 많은 수의 문서가 검색되었음에도 불구하고 적합문서의 수가 매우 적었고, 평균 정확률과 n-순위 정확률이 가장 낮게 나타났다. 이는 실험 집단 내 앵커텍스트가 웹 문서의 내용을 표현하기에 부적합한 정보를 포함하는 경우가 많기 때문인 것으로 해석된다. 실험집단의 앵커텍스트를 살펴보면, 다음 예와 같이 Home, Back 등과 같은 웹 항해를 위한 링크이거나, 앵커텍스트의 내용이 링크 페이지의 주소나, 페이지번호, 일련번호 등과 같이 실제 질의의 내용과는 관련이 없는 경우가 종종 나타났다.

〈a href="index.html"〉[HOME]〈/a〉
〈a href="/meeting.html"〉Back〈/a〉
〈A HREF="dglyrics.html"〉dglyrics.html〈/A〉
〈a href="images/condchg.gif"〉〈B〉Figure 1〈/B〉〈/a〉
〈A HREF="../textdoc/features.html"〉page 53〈/A〉

마지막으로 URL 정보의 경우 평균 정확률은 3.3%로서 가장 높게 나타났으며, n-순위 정확률은 앵커텍스트와 유사한 수준으로서 내용정보에 비해 50% 가량 낮게 나타났다. 앵커텍스트의 경우 10개 질의 중 8개 질의에서 적합문서가 검색되었으며 정확률도 질의에 따라 큰 차이를 보이지 않았지만 URL의 경우에는 10개 질의 중 6개의 질의가 적합문서를 검색하지 못한 반면, 4개의 질의에서는 0.35로 매우 높은 정확률을 보여 질의에 따른 성능차이가 다른 문서표현 방법에 비해 큰 것으로 나타났다.

이는 일부 웹 문서의 저자들이 URL을 표현할 때 내용을 집약적으로 나타내기 위해 URL 표현에 신중을 기한 것으로 해석된다. 덧붙여 URL 정보는 유사한 정확률 수준을 보인 앵커텍스트보다 수집이 용이할 뿐만 아니라 데이터의 양도 적어 간단하게 이용할 수 있으므로 웹 문서의 표현에서 내용정보와 링크 정보만으로 검색되지 않는 웹 문서를 검색하기 위한 방법으로 고려될 수 있을 것이다.

3.2 문서표현 방법을 결합한 검색결과

3.2.1 두 가지 문서표현 방법의 결합

실험에서 사용된 내용, 앵커텍스트, URL 등의 세 가지 표현방법에 대해 표현 방법을 두 가지씩 결합하면 내용-앵커텍스트(C_A), 내용-URL(C_U), 앵커텍스트-URL(A_U)의 결합이 있게 된다. 〈표 5〉는 두 가지 표현방법의 결합에서 나타난 검색문서 수, 검색된 문서 중 적합문서 수, 새로 검색된 적합문서 수, 중복문서 수, 중복문서 중 적합문서 수를 보여준다. 여기서 새로 검색된 적합문서 수는 앞쪽에 표기된 문서 표현 방법을 기준으로 산

출되는데, 예를 들어 결합방식이 C_A일 경우 내용정보를 기준으로 앵커텍스트가 추가되면서 새로 검색된 적합문서 수를 의미한다.

내용-앵커텍스트, 내용-URL, URL-앵커텍스트 결합결과 검색문서 수의 평균은 각각 2248.5, 521.6, 2022.2개로 나타났으며, 검색된 문서 중 적합문서의 수는 평균 10.4, 9.3, 6.2개로 검색된 문서 중의 적합문서 수는 1% 미만으로 적게 나타났다. 또한 중복문서의 비율은 모두 10% 미만으로 나타났으며, 결합방식의 첫 번째 알파벳을 기준으로 할 때 새로 검색된 적합문서의 수는 평균 1.1, 0, 0.9개였다.

중복문서에 대한 적합문서의 비율은 각각 41.35%, 26.88%, 24.19%로 내용과 앵커텍스트를 결합하였을 때 중복률이 높았고 링크와 URL 정보의 중복률이 가장 낮게 나타났다.

〈표 6〉은 두 가지 표현방법을 결합했을 때의 평균 n-순위 정확률과 성능향상률을 나타낸 것이다. 여기에서 표현방법의 결합 시 사용한 〈공식 3〉의 가중치 w_i는 모두 1로 하였다. 성능향상률은 결합에 사용된 표현방법 중 단일문서 표현방법에서 가장 성능이 우수했던 표현방법과 비교하여 산출하였다.

두 가지 표현방법 결합의 경우 앵커텍스트-URL 결합을 빼고는 모두 성능의 저하를 가져왔다. 이러한 결과는 결합의 성패여부가 중복문서 수, 중복률에 따라 달라질 수 있음을 의미한다. 중복문서 수가 평균 188개로 상대적으로 많고, 중복률도 41.35%로 상대적으로 높은 내용-앵커텍스트 결합에서는 평균 n-순위 정확률에서 단일 표현 방법보다 비교적

〈표 5〉 두 가지 표현방법 결합 실험 결과

결합방식	검색문서 수	검색된 문서 중 적합문서 수(%)	새로 검색된 적합문서 수	중복문서 수 (%)	중복된 문서 중 적합문서 수(%)
C_A	2248.5	10.4(0.5%)	1.1	188(8.3%)	4.3(41.35%)
C_U	521.6	9.3(1.8%)	0	46.6(8.93%)	2.5(26.88%)
A_U	2022.2	6.2(0.3%)	0.9	36.1(1.81%)	1.5(24.19%)

〈표 6〉 두 가지 표현방법의 결합결과 평균 n-순위 정확률과 성능향상률
(동일한 가중치 사용)

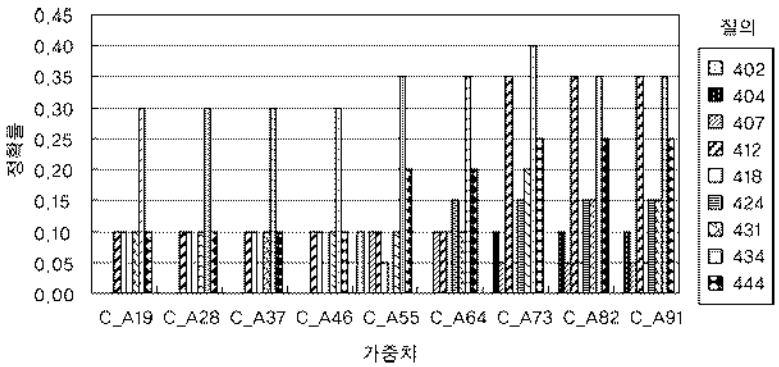
결합방식	비교대상	P(5)	성능향상률 (%)	P(10)	성능향상률 (%)	P(20)	성능향상률 (%)
C_A	C	0.2	-12	0.15	-10	0.1	-9
C_U	C	0.18	-14	0.17	-8	0.125	-6.5
A_U	U	0.14	0	0.16	+1	0.145	+6

큰 성능하락을 가져왔다. 또한 중복문서 수가 평균 46.6개로 많지 않고, 중복률이 26.88%로 높은 내용-URL 결합에서는 평균 n-순위 정확률이 다소 하락하는 것에 그쳤다. 반면, 중복문서 수도 적고(36.1개) 중복률도 낮은(24.19%) 앵커텍스트-URL 결합에서는 동일하거나 다소 우수한 성능을 보였다.

데이터 결합 성능은 결합가중치에 따라 달라질 수 있으므로 〈공식 3〉의 결합 가중치 w_i값을 0.1~0.9까지 0.1 단위씩 증가시켜 두 가지 표현방법을 결합하는 실험을 수행한 결과 세 경우(C_A, A_U) 모두 두 가지 표현 방법 중 성능이 좋은 내용기반 표현에 더 높은 가중치C_U, 를 부여할 경우 결합성능이 높게 나타났다.

〈그림 3〉은 내용-앵커텍스트 결합 시 질의별 성능을 보여 준다. 그림에서 C_A1,9는 내용 정보(C)에 가중치 0.1을, 앵커텍스트(A)에 가중치 0.9를 부여한 경우를 나타낸다. 이 경우 문서표현을 결합한 신념 값 bel(I|D_j)를 산출하는 〈공식 3〉에서 내용정보(R_1)에 대한 신념의 결합가중치 w_1은 0.1, 앵커텍스트(R_2)에 대한 신념의 가중치 w_2는 0.9가 된다. 내용정보와 앵커텍스트의 결합에서는 내용 정보에 0.7, 앵커텍스트에 0.3의 가중치를 부여하였을 때 가장 좋은 성능을 보였다.

〈표 7〉은 두 가지 표현방법의 결합(C_A, C_U, A_U)에서 각각 최적의 결합 가중치를 사용했을 경우의 검색성능을 보여 준다. 〈표 6〉과 마찬가지로 평균 n-순위 정확률과 성능향상률로 평가하였으며, 성능향상률은 결합에 사용된 문서표현 방법 중 단일 문서표현 방법에서 우수한 성능을 보인 표현방법과 비교하였다. 최적의 가중치를 사용하였을 경우 내용-앵커텍스트 결합에서는 내용정보만을 사용하였을 때와 성능차이가 없었고, 동일한 가중치에 비해서는 훨씬 좋은 성능을 보였다. 내용-URL 결합은 동일한 가중치를 사용하였을 때와 유사한 결과를 보였고, 앵커텍스트-URL 결합은 동일한 가중치에 비해 P(5),



〈그림 3〉 내용-앵커텍스트 결합시 결합가중치에 따른 질의별 검색성능 비교

〈표 7〉 두 가지 표현방법의 결합결과 평균 n-순위 정확률과 성능향상률
(최적 가중치 사용)

결합방식	비교대상	P(5)	성능향상률 (%)	P(10)	성능향상률 (%)	P(20)	성능향상률 (%)
C_A	C	0.32	0	0.25	0	0.19	0
C_U	C	0.18	-14	0.17	-8	0.16	-3
A_U	U	0.20	+6	0.12	+3	0.05	-3.5

P(10)에서 좋은 성능을 보였다.

3.2.2 내용-앵커텍스트-URL 표현방법의 결합

내용, 앵커텍스트, URL을 모두 결합한 실험에서 검색된 문서 수는 평균 2274.9개, 검색문서 중 적합문서 수는 평균 10.4개, 중복문서 수는 평균 24.5개, 중복문서 중 적합문서 수는 평균 1.0개로 나타났다. 내용-앵커텍스트 결합과 비교 하면 검색문서 수에서는 내용-앵커텍스트의 평균 검색문서 수인 2248.5개와 비슷하였고, 검색문서 중 적합문서의 수는 동일하게 나타났다. 그러나 세 표현의 결합 결과 중복 검색된 문서 중 적합문서는 거의 없었다.

내용-앵커텍스트-URL 결합 시에도 결합 가중치를 다르게 부여하여 실험을 수행하였다. 앞의 실험에서 내용정보의 가중치를 높이는 것이 결합 시 검색성능을 높여주는 것으로 나타났기 때문에 세 방법에 가중치를 동일하게 주었을 경우의 가중치 0.33보다 큰 값인 0.4부터 0.1씩 증가시키면서 내용정보의 가중치를 부여하였다.

내용정보의 가중치 값을 정한 다음 앵커텍스트와 URL은 두 방법의 성능이 비슷한 것으로 나타났기 때문에 남은 가중치 값을 이분하여 동일한 가중치를 부여하였다. 예컨대 내용정보가 0.4의 가중치를 갖는 경우 앵커텍스트와 URL은 각각 0.3의 가중치 값을 갖는다. 〈표 8〉에서 가중치 결합을 C_A_U6,2와 같이 표현하였는데, 이는 내용정보에 0.6의 가중치를 부여하고 나머지 두 표현방법에 대해 0.2

〈표 8〉 내용-앵커텍스트-URL 결합 시 평균 n-순위 정확률

	C_A_U4,3	C_A_U5,2.5	C_A_U6,2	C_A_U7,1.5	C_A_U8,1
P(5)	0.20	0.23	0.30	0.34	0.34
P(10)	0.19	0.18	0.24	0.24	0.26
P(20)	0.11	0.13	0.16	0.17	0.17
평 균	0.165	0.178	0.233	0.25	0.257

씩의 가중치를 부여한다는 의미이다.

〈표 8〉은 단일 표현으로 가장 성능이 좋은 내용 정보에 높은 가중치 값을 부여한 경우 성능이 좋게 나타남을 보여 준다. 특히 내용 정보의 가중치에 0.8을 부여하고 나머지 표현 방법에 0.1씩의 가중치를 부여한 경우가 성능이 가장 좋았다.

3.2.3 링크분석을 이용한 검색확장

링크를 이용한 검색에서는 웹 문서 간 직접적으로 연결된 문서들을 파악하여 검색에 활용하는 직접링크 분석, 2개의 웹 문서가 제3의 웹 문서에 의해 동시에 링크된 상태를 이용하는 동시인용 기법, 2개의 웹 문서가 동일한 웹 문서를 공통적으로 링크하고 있는 상태를 이용하는 서지결합 기법 등 세 가지 방법에 의해 검색실험을 수행하였다.

검색실험에서는 성능이 가장 좋은 내용정보에 의한 검색결과를 이용하여 적합문서들의 링크정보를 추출하였다. 이때 링크하고 있는 웹 문서 파일은 본 실험에서 사용한 실험집단(WT2g_sub)으로부터 확보하였고, 링크되고 있는 웹 문서 파일은 실험집단 전체(WT2g)를 대상으로 하였다.

직접링크를 이용하여 검색을 수행한 결과 검색된 문서 수와 적합문서 수는 적었지만 다른 표현방법에 의해 검색되지 않은 새로운 문서들이 검색되었다. 각 질의에 대해 검색된 문서의 수는 평균 34.8개였으며, 이 중 새로운 검색문서는 평균 18개였다. 그러나 검색된 문서 중 적합문서의 수는 평균 1개로서 매우 낮게 나타났다.

이러한 결과는 서지결합과 동시인용 기법을 이용한 실험에서도 유사하게 나타났다. 서지결합의 경우 검색문서 수와 새로운 검색문서 수가 각각 18개와 17개이며, 검색된 문서 중 적합문서 수는 0.8개로 직접링크를 사용한 경우와 별 차이가 없었다. 반면 동시인용의 경우 검색된 문서의 수가 1,166개로서 엄청나게 증가한데 비해 검색된 적합문서의 수는 21개로서 매우 적었다. 실험결과가 이처럼 나타난 것은 적합문서가 웹 사이트의 초기화면일 경우, 해당 사이트에 포함되어 있는 웹 페이지의 대부분이 초기화면을 가리킴으로써 동시인용에 따른 검색문서 수가 증가한 것으로 해석된다.

하지만 링크분석 기법을 이용하였을 때와 단일 문서표현 방법을 비교하여 보면, 동시인용의 경우 검색된 문서 수가 앵커텍스트를 이용한 검색결과와 유사하게 나타났지만 앵커텍스트를 이용한 검색문서 중 적합문서의 비율이 0.3%였던데 반해, 동시인용을 이용한 경우는 1162.7개의 새로운 문서를 검색하였고, 이중 21개가 적합문서였으므로 적합문서 비율이 약 1.8%로 증가하였다. 또한 직접링크와 동시인용 분석의 경우 적합문서의 비율이 각각 5.46%, 4.82%를 나타냄으로써 새로운 적합문서를 찾는 데 있어서 두 기법이 단일 문서표현 방법보다 유용할 수 있음을 알 수 있었다.

4. 결 론

이 연구에서는 데이터 결합이 웹 문서의 검색성능에 어떤 영향을 미치는 지를 실험을 통해 분석하였다. 실험결과 웹 문서의 경우 데이터 결합이 검색성능의 향상에 별로 기여하지 못하는 것으로 나타났다. 그러나 동일한 정보요구에 대해 문서의 상이한 표현을 이용하여 중복되지 않는 문서를 많이 검색할 경우에는 데이터 결합이 검색성능의 향상을 가져올 수 있음을 확인하였다. 구체적인 실험결과는 다음과 같다.

첫째, 단일문서 표현에서 내용정보를 이용한 문서표현 방법이 가장 성능이 좋은 것으로 나타났고, 그 다음으로 URL, 앵커텍스트의 순이었다. 특히, 내용정보는 다른 표현방법에 비해 2배 이상 좋은 검색성능을 보여 내용정보가 단일 문서표현에서는 가장 적합한 방법이라는 사실을 확인하였다.

둘째, 내용정보를 표현한 두 가지 방식에서 〈title〉 태그만을 이용하는 것보다 〈title〉, 〈h〉, 〈font〉 태그를 함께 이용하는 것이 n-순위 정확률에서는 다소 향상된 결과를 보였다. 이는 웹 문서의 전문을 사용하기 어려운 환경에서는 〈h〉 태그와 〈font〉 태그를 〈title〉 태그와 함께 사용하였을 때 웹 문서의 내용을 더 잘 표현할 수 있음을 시사한다.

셋째, 내용-앵커텍스트, 내용-URL, 앵커텍스트-URL 등의 두 가지 문서표현 방법을 결합한 결과, 검색된 문서의 중복률이 비교적 높은 내용 표현방법과 그 외의 두 가지 문서표현방법을 결합하였을 때는 검색성능이 오히려 낮아지는 결과를 보였다. 그러나 상대적으로 중복률이 낮은 앵커텍스트와 URL을 결합하였을 때는 검색성능이 다소 향상되는 결과를 보였다.

넷째, 내용-앵커텍스트-URL을 모두 결합한 경우에는 검색성능이 오히려 저하되는 것으로 나타났다. 그것은 내용정보의 검색성능이 다른 표현방법에 비해 현저히 우수하기 때문에 성능이 낮은 다른 표현과 결합은 검색성능의 저하로 나타난 것으로 보인다.

다섯째, 링크를 이용한 확장검색은 다른 표현방법을 이용한 검색결과를 개선하지 못한 것으로 나타났다.

결론적으로 웹 문서 검색에 있어서 문서표현 방법의 결합은 단일 문서표현 방법에 비해 뚜렷한 성능향상을 가져오지는 않았지만, 여러 다른 표현방법의 결합이 상대적으로 많은 적합문서를 검색해 주므로 포괄적인 정보검색에 있어서는 유용할 수 있을 것으로 보인다.

이 연구는 실험 시스템의 한계로 인해 TREC의 WT2g 실험집단 중 일부만을 실험에 사용하였고, 링크분석에서는 실험집단에 포함된 웹 문서간 링크 정보만을 이용하였기 때문에 실험결과가 웹 검색 환경을 충분히 반영하지 못했다는 제한점을 가지고 있다.

참고문헌

안동언, 강인호. 2002. 웹 정보검색 시스템의 문서 순위 결정. 「정보관리연구」, 34(2): 55-66.

전상우. 2005. 「색인 결합을 이용한 검색 성능 향상에 관한 실험적 연구」. 석사학위논문, 연세대학교 대학원, 문헌정보학과.

정영미. 2005. 「정보검색연구」. 서울: 구미무역(주) 출판부.

최성환. 2001. 「용어 가중치 결합의 검색 효율성에 관한 연구」. 석사학위논문, 연세대학교 대학원, 문헌정보학과.

Allan, J., Callan, J., Sanderson, M., Xu, J., and Wegmann, S. 1998. "INQUERY and TREC-7". *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*. NIST Special Publication 500-242. [cited 2006.5]. 〈http://trec.nist.gov/pubs/trec7/t7_proceedings.html〉.

Amitay, E. 1998. "Using common hypertext links to identify the best phrasal description of target Web documents". *Proceedings of the SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web*.

Belkin, N. J., Coll, C., Croft, W. B., and Callan, J. P. 1993. "The effect of multiple query representations on information retrieval performance". *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 339-346.

Belkin, N. J., Kantor, P., Fox, E. A., and Shaw, J. A.. 1995. "Combining the evidence of multiple query representations for information retrieval". *Information Processing & Management*, 31(3): 431-448.

Bharat, K., and Henzinger, R. 1998. "Improved algorithms for topic distillation in a hyperlinked environment". *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 64-71.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., and Rajagopalan, S. 1998. "Automatic resource list compilation by analysing hyperlink structure and associated text". *Proceedings of the 7th International World Wide Web conference*.

Croft, W. B. 2000. *Combining approaches to information retrieval*. In Croft, W.B., ed. Advances in Information Retrieval. Boston: Kluwer Academic Publishers.

Davidson, B. D. 2000. "Topical locality in the Web". *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 272- 279.

Fox, E. A., and Shaw, J. A., 1993. "Combination of multiple searches". *Proceedings of the Second Text Retrieval Conference(TREC-2)*. NIST Special Publication 500-215. [cited 2006.5]. <http://trec.nist.gov/pubs/trec2/t2_proceedings.html>.

_____. 1994. "Combination of multiple searches". *Proceedings of the Third Text Retrieval Conference(TREC-3)*. NIST Special Publication 500-225. [cited 2006.5]. <http://trec.nist.gov/pubs/trec3/t3_proceedings.html>.

Hawking, D. Voorhees, E., Craswell, N., and Bailey, P. 1999. "Overview of the TREC-8 web track". *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. NIST Special Publication 500-246. [cited 2006.5]. <http://trec.nist.gov/pubs/trec8/t8_proceedings.html>.

Hawking, D. 2000. "Overview of the TREC-9 web track". *Proceedings of the Ninth Text Retrieval Conference(TREC-9)*. NIST Special Publication 500-249. [cited 2006.5]. <http://trec.nist.gov/pubs/trec9/t9_proceedings.html>.

Hawking, D., and Craswell, N., 2001. "Overview of the TREC-10 web track". *Proceedings of the Tenth Text Retrieval Conference(TREC-10)*. NIST Special Publication 500-250. [cited 2006.5]. <http://trec.nist.gov/pubs/trec10/t10_proceedings.html>.

Jansen, B. J. and Spink, A. 2005. "An analysis of Web searching by European alltheWeb.com users". *Information Processing and Management*, 41: 361-381.

Jansen, B. J., Spink, A., and Saracevic, T. 2000. "Real life, real users, and real needs: a study and analysis of user queries on the Web". *Information Processing and Management*, 36: 207-227.

Kang, I. H., and Kim, G. C., 2003. "Query type classification for web document retrieval". *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*,

64-71.

Katzer, J., Tessier J., Frakes, W., and DasGupta, P. 1983. "A study of the overlap among document representations". *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 106-114.

Lee, J. H. 1995. "Combining multiple evidence from different properties of weighting schemes". *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 180-188.

_____. 1997. "Analyses of multiple evidence combination". *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, 267-276.

McGill, M. K. and Noreault, T. 1979. *An investigation of factors affecting document ranking by information retrieval system*. Technical report, School of Information Studies, Syracuse University.

Missingham, R. 1996. "Indexing the Internet: pinning jelly to the wall?". *Library Automated Systems Information Exchange*, 27(3): 32-42.

Ogilvie, P., and Callan, J. 2003. "Combining document representations for known-item search". *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 27-34.

Rajashekar, T. B., and Croft, W. B. 1995. "Combining automatic and manual index representations in probabilistic retrieval". *Journal of the American Society for Information Science*, 46(4): 272-283.

Salton, G., Allan, J., and Buckley, C. 1993. "Approaches to passage retrieval in full text information systems". *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 49-58.

Thelwall, M. and Wilkinson, D. 2004. "Finding similar academic Web sites with links, Bibliometric Couplings and Colinks". *Information Processing and Management*, 40: 515-526.

Tsikrika, T. and Lalmas, M. 2002. "Combining Web document representations in a Bayesian inference network model using link & content-based evidence".

Proceedings of the 24th European Colloquium on Information Retrieval Research(ECIR 2002), 53-72.

_____. 2004. "Combining evidence for Web retrieval using the inference network model: an experimental study", *Information Processing and Management*, 40(5): 751-772.

Turtle, H. R. 1990. Inference Networks for Document Retrieval. Ph.D. diss., University of Massachusetts, Amherst.

Turtle, H. R. and Croft, W. B. 1991. "Evaluation of an inference network-based retrieval model", *ACM Transactions on Information System*, 9(3): 187-222.

Vogt, C. C., and Cottrell, G. W. 1998. "Predicting the performance of linearly combined IR systems", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 190-196.

Westerveld, T. Kraaij, W., and Hiemstra, D. 2001. "Retrieving web pages using content, links, urls and anchor", *Proceedings of the Tenth Text Retrieval Conference(TREC-10)*. NIST Special Publication 500-250. [cited 2006.5]. 〈http://trec.nist.gov/pubs/trec10/t10_proceedings.html〉.

Yang, K., 2001. "Combining text- and link-based retrieval methods for Web IR", *Proceedings of the Tenth Text Retrieval Conference(TREC-10)*. NIST Special Publication 500-250. [cited 2006.5]. 〈http://trec.nist.gov/pubs/trec10/t10_proceedings.html〉.