Link와 Clustering을 이용한 적극적 문서 수집 기법

김원우 변영태
홍익대학교 컴퓨터공학과

Greedy Document Gathering Method Using Links and Clustering

Won Woo Kim Young Tae Byun
Dept. of Computer Science, Hong-Ik University

요약

특정 영역에 대해 사용자에게 관련 정보를 제공해 주는 서비스를 하는 정보 에이전트를 개발 중이다. 정보 에이전트는 사용자 질의 처리를 맡은 Agent Manager와 지식베이스를 관리하는 KB Manager, 그리고 Web으로부터 해당 영역의 관련 문서를 끌어오는 Web Manager로 구성되어 있다. Web Manager는 방문할 URL을 수집하고, 이들 문서에 대한 관련 평가와 Indexing을 수행한다. Web Manager는 검색 엔진을 이용하거나, 방문한 문서의 link를 이용하여 URL을 수집하는데, 이러한 URL수집기법은 많은 관련 문서를 놓치는 문제점이 있다. 이 문제점을 해결하기 위해서 해당 영역과 관련된 Site들을 대상으로 Link를 이용해 문서들을 모아와, 문서들을 TAG들의 패턴으로 얻어낸 문서 형식을 이용해 Clustering하여 관련 문서들의 Group을 찾아내는 적극적 문서 수집 기법을 제안한다. 실험 결과, Link와 Clustering을 이용할 경우 기존보다 효과적으로 관련 문서를 많이 수집할 수 있음을 알 수 있다.

1. 서론

정보 에이전트는 특정 영역에 대한 문서를 Database에 저장해 놓았다가, 사용자가 원하는 정보를 담고 있는 관련 문서를 제공해 주는 서비스를 수행한다. 따라서 정보 에이전트가 특정 영역과 관련된 문서를 많이 가지고 있어야 좋은 서비스를 해줄 수가 있다. 즉, 정보 에이전트는 좋은 문서를 많이 수집할 수 있어야 한다[1].

정보 에이전트의 Database를 분석한 결과, 관련 문서들이 있는 Site에서 방문하지 않은 관련 문서들이 더 있다는 것을 발견하였다. 그 문서들은 관련 문서들과 비슷한 문서 형식으로 작성되어 있었다. 이런 특성을 잘 이용하면, 관련 문서들을 효과적으로 수집할 수 있을 것이다.

연구의 기본 정책은 기존의 관련 문서들이 있는 Site에 숨겨져 있는 문서들을 찾아내는 것이다. 이를 위해 Link를 효과적으로 이용한다. 내부로의 Link는 방문하는 Site내의 문서를 가지고 오는데 이용한다. 반면 외부로의 Link는 Site의 방문이 끝난 후 다음에 방문할 Site를 선택하는 정보로 또는 새로운 관련 Site를 찾는데 이용한다. 관련 문서들을 문서 형식을 이용한 Clustering을 통해 Group단위로 찾고자 한다.

본 논문의 구성은 다음과 같다. 제 2 절에서는 기존의 정보 에이전트에 대해 살펴본다. 제 3 절과 제 4 절에서는 적극적 문서 수집 기법에 대한 내용과 구현 내역에 대해 설명하며, 제 5 절에서 실험에 관하여 설명하고 마지막으로 제 6 절에서 결론을 맺는다.

□□□□□□□□□□□□□□□□□□□□□□□□□□
본 연구는 뇌과학 연구 사업의 지원으로 진행 되었음.

2. 정보 에이전트

정보 에이전트는 Web을 기반으로 특정 영역에 대해 사용자가 원하는 문서를 제공해 주는 System을 말한다. 정보 에이전트의 개발을 위해 특정 영역으로는 동물 영역을 선택하였다[1]. 정보 에이전트는 사용자의 질의를 받아서 그 질의에 적합한 문서를 찾아주는 Agent Manager와 에이전트의 지시베이스를 관리하는 KB Manager, Web으로부터 특정 영역에 대한 정보를 담고 있는 관련 문서를 끌어오는 일을 맡은 Web Manager로 구성된다. Web Manager의 URL Robot은 방문할 문서의 URL들을 수집한다. 수집된 URL들을 Web Manager의 Indexer가 방문해서 특정 영역과 관련된 문서라면 정보 에이전트의 Database에 저장해 놓는다[2].

정보 에이전트는 이렇게 구축된 Database에서 사용자로부터 받은 질의에 대한 답을 담고 있는 적절한 문서들을 찾아 준다. 따라서 Database에 특정 영역에 관련된 문서의 수가 적거나, 적합한(전문) 문서의 수가 적을 경우에는 질의에 대한 적절한 대답을 해 줄 수가 없다. 그러므로 정보 에이전트는 관련 문서들을 Database에 가능한 많이 저장해 놓아야 효율적인 서비스를 할 수가 있다.

관련 문서를 모으기 위해 Web Manager의 URL Robot은 일반 검색 엔진에 정보 에이전트의 KB를 적절히 조합한 Keyword들을 보내어 나온 결과 URL들을 수집한다. 다른 방식으론 방문한 관련 문서들에 포함되어 있는 Link들로부터 URL들을 수집한다.

위에서 설명한 URL수집기법으론 많은 관련 문서를

놓치는 문제점이 있다. 일반 검색 엔진을 이용할 경우에는 검색엔진이 범용이기 때문에 여러 영역에서 고루 이용되는 키워드로 검색된 결과 문서들 중 극히 일부만이 관련된 문서들이다.(예 : 동물 영역의 tiger – tiger mask, tiger woods). 또한, 정보 에이전트의 KB가 미흡하다면, KB에 없는 Data가 포함된 문서들은 찾아올 수가 없다. 또, Link를 이용하는 경우에는 지나친 검색 범위의 확장을 막기 위해 Link를 쫓아가는 깊이의 제한을 두기 때문에 깊이 이상에 있는 좋은 문서들을 놓칠 위험이 있다.

3. 문서의 수집

특정 영역과 관련된 일련의 Site들을 알 수 있다면, 정보 에이전트는 Site를 직접 방문해 관련 문서들을 확실하게 수집해 올 수 있을 것이다. 먼저 Site 정보를 효과적으로 관리하기 위해 Site Table를 구성한다. Site Table의 여러 Site중에 방문할 가치가 높은 Site를 선택한다. 선택된 Site에 포함되어 있는 모든 문서들을 Link를 이용해 가져온다. 모아진 Link들은 내부로의 Link인지, 외부로의 Link인지에 따라 적절히 이용된다.

3.1 Site Table

방문할 Site Selection에 사용될 Site가 가진 관련 문서의 수, 다른 Site로부터의 Link되어진 수 등의 정보를 관리하는 Site Table을 구성한다. Site Table은 다음과 같이 생성하고 관리한다.

생성

정보 에이전트에는 이미 관련 문서로 판단된 문서들이 Database에 있다. 여기에는 관련 문서의 URL과 Title, Description 등의 정보가 있다. 관련 문서의 URL들을 Site별로 몇 개의 관련 문서를 가지고 있는지 정리하면, 어떤 Site에 관련 문서들이 얼마나 있는지 알 수가 있다. 그런 Site에는 아직 방문하지 않은 관련 문서들이 더 있을 가능성이 높다.

SITE	관련문서수	Link된수
http://animaldiversity.ummz.umich.edu/	5	0
http://www.parks.tas.gov.au/	4	0
http://www.animalinfo.org/	4	0
...	...	

(표 1) 생성된 Site Table의 예

관리

방문한 Site에서 찾아진 외부로의 Link들을 Site Table에 추가한다. 이렇게 하면 Site Table에 없는 새로운 관련 Site들을 모을 수도 있고, 더 나은 Site를 선택할 수 있는 정보도 모을 수 있다.

3.2 Site Selection

제안하는 기법의 기본 정책은 한 Site에 포함되어 있는 모든 문서들을 가지고 와서 Clustering해 보자는 것이다. 그러므로 System은 여러 Site들 중에 방문할 Site의 선택을 효율적으로 해야 한다. 방문할 Site를 선택하기 위해서는 VV(Visiting Value: Site의 방문 가치)를 계산한다. VV 계산에는 다음의 두 인수들을 이용한다.

RPN : Relevant Pages Number

관련 문서가 많이 포함되어 있는 Site일수록 좋은 Site라고 여겨질 수 있다. 따라서 각 Site에 관련 문서가 몇 개씩 있는지를 조사하여, RPN이 높은 순으로 Site를 선택 하고자 한다.

LN : Linked Number

LN은 어떤 Site가 같은 영역에 관련된 다른 Site들로부터 Link되어진 수이다. 어떤 영역에 대해 한 Site가 다른 Site들로부터 link되어진 수가 많을수록 그 Site가 그 영역에서 좋은 Site라고 여겨질 수 있다[3]. 따라서 LN이 높은 순으로 Site를 선택하고자 한다.

VV는 다음과 같은 방식으로 계산할 수 있다. 계산된 VV가 높은 순으로 Site를 선택한다.

$VV_{weighted} = w_1 \times RPN + w_2 \times LN$

$VV_{sum} = RPN + LN$

$VV_{mul} = RPN \times LN$

SITE	RPN	LN	$VV_{weighted}$	VV_{sum}	VV_{mul}
A	5	5	15	10	25
B	4	6	14	10	24
C	3	7	13	10	21

(표 2) VV의 계산 예

($VV_{weighted}$: w_1=1, w_2=2)

실험에서는 $VV_{weighted}$를 사용하였다. $VV_{weighted}$를 이용하면 표 1에서 http://animaldiversity.ummz.umich.edu/가 선택되게 된다.

3.3 Link

Site가 결정되면, 홈페이지부터 시작하여 문서들의 Link들을 쫓아가 Site에 속해 있는 모든 문서들을 수집한다. Web Site의 구조 분석을 위해 Web 문서들을 계층 구조나 그래프로 표현하기도 한다[4]. 그렇지만, 제시하는 기법에서는 구조 분석을 도입해 문서를 분류하지 않는다. 문서로부터는 <A> 혹은 <Frame>에 속한 Link를 추출한다.

내부로의 Link

추출되어진 Link중에 방문하고 있는 Site로의 Link는 Site에 속한 모든 문서들을 가지고 오는데 이용한다. 추출되어진 Link들 중에 같은 Site로의 Link만을 쫓아가면 그 Site에 속한 모든 문서들을 가지고 올 수 있다.

외부로의 Link

추출되어진 Link중에 다른 Site로의 Link는 다음 방문할 Site의 결정에 이용되는 LN에 추가된다. Site Table에 없는 Link는 새로운 관련 Site로 Site Table에 추가된다. 외부로의 Link를 List(외부로의 Link List)에 저장한다.

3.4 외부로의 Link List

외부로의 Link List에 모아진 외부 Site들의 정보를 기존의 Site Table에 추가해 준다. 만약 외부로의 Link List가 다음과 같다면, Site Table은 표와 같이 변할 것이다.

http://www.animalinfo.org/
http://animaldiversity.ummz.umich.edu/
http://biology.usgs.gov/

SITE	RPN	LN
ttp://animaldiversity.ummz.umich.edu/	5	1
ttp://www.parks.tas.gov.au/	4	0
ttp://www.animalinfo.org/	4	1
ttp://biology.usgs.gov/	0	1
...	...	...

(표 3) 새로워진 Site Table

이제 System은 http://www.animalinfo.org/을 다음에 방문할 것이다.

4. Clustering

Clustering은 World wide web상에서 문서들을 분류하는데 사용되어 왔다. Clustering에는 여러 가지 기법들(graph partitioning등[5])이 이용되어 왔다. 문서 Clustering의 feature로는 주로 word나 text가 사용되어 왔다.

대규모 정보 Web Site는 비슷한 문서 형식으로 양질의 정보들을 제공하고 있다. 이를 이용해 Clustering을 하면 보다 쉽게 관련 문서들을 찾아 올 수 있다. 기존처럼 문서들을 개별적으로 방문해 내용을 가지고 관련 여부를 평가하는 대신에 문서의 형식을 이용해 우선 비슷한 문서들을 Group으로 묶어 놓고, 그 다음에 좋은 Group들을 선택한다.

Link들을 쫓아 문서들을 방문하면서 동시에 문서들의 형식을 Clustering에 이용될 Matrix로 표현해 놓는다. Site 방문이 끝나고 나면, Matrix들을 이용해 문서들을 Clustering한다. 얻어진 여러 결과 Group들 중에 관련 문서가 많이 포함되어 있는 좋은 Group을 선택해 정보 에이전트의 Database에 저장한다.

4.1 TAG Matrix

TAG가 이용된 Pattern을 문서마다 TAG Matrix의 형태로 문서 형식을 구성한다.

TAG

문서 형식은 TAG에 의해서 결정된다. TAG들은 보편적으로 이용되는 것들 중 '여는 TAG'들만을 이용한다. 유의할 점은 TAG<a>를 제외시킨 것이다. TAG<a>는 자주 이용되고, 이용되는 장소가 일정치 않아 제외시킨다.

p,li,td,b,br,small,tr,img,font,em,table,hr,h3,i, blockquote,ul,h2,sup,cite,area,strong,u,nobr,h4, center,h1,style,map,script,dd,dt,meta,address,tt, input,link,option,caption,form,div,ol,dl,dir,select,th

(표 4) TAG Array (45개)
(효율적으로 이용하기 위해서 실제로 Web에서 많이 이용되는 순으로 Sorting)

TI(TAG) = TAG Array에서 TAG의 Index
(예 : TI(p) = 0, TI(br) = 4)

TP : TAG Pattern

어떤 문서를 대표하는 문서 형식의 특징은 문서에 이용된 TAG들의 순서에 의해 결정된다. 이를 TAG Pattern이라 한다. TP는 TAG Array에 속한 TAG들만을 대상으로 구성한다. 문서 k의 TP는 다음과 같이 표현된다.

$$TP_k = \{TAG_0, TAG_1, ..., TAG_n\}$$

TAG Matrix

TP를 Clustering에 이용하기 위해서 TAG Matrix의 형태로 변환한다. 먼저 표4의 TAG들을 이용해 표5와 같은 Symmetric Matrix(2차원 배열)을 만든다.

TP의 TAG_x의 x가 0에서 n-1까지 Matrix $[TI(TAG_x)][TI(TAG_{x+1})]$에 Boolean True를 표시해 준다.

	<P>	<LI>	<TD>		<TH>
<P>					
<LI>	True				
<TD>		True			
<TH>					

(표 5) TAG Matrix (45 × 45)
(예 : TP={<TD>,<LI>,<P>})

문서 d의 Matrix는 다음과 같이 나타낸다.

M_d : 문서 d의 Matrix

t개의 문서로 이루어진 Site는 다음과 같이 나타낼 수 있다.

$$Site = \{M_0, M_1, \ldots, M_{t-1}\}$$

4.2 Clustering

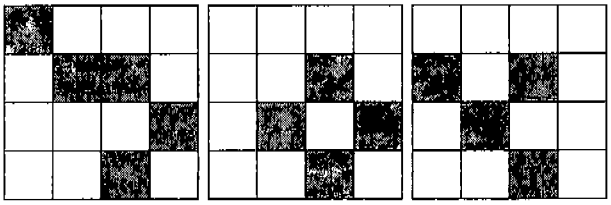
Clustering은 Matrix간의 유사성을 비교하여, 비슷한 Matrix들을 Group으로 묶는다.

Distance Function

Matrix간의 거리를 계산해서 가까운 문서들끼리 묶는 간단한 방법을 이용한다. 이에 이용되는 Distance Function은 다음과 같다.

$$Dist(M_i, M_j) = SAME(M_i, M_j) / TRUE(M_i)$$

$SAME(M_i, M_j)$: M_i와 M_j간에 일치하는 Boolean True의 수
$TRUE(M_i)$: M_i의 Boolean True의 수



M_1 M_2 M_3

(그림 1) TAG Matrix들

$Dist(M_1,M_2) = 3/5 = 0.6$ $Dist(M_1,M_3) = 2/5 = 0.4$

Clustering - Matrix

아직 Group에 속하지 않은 문서 M과 이미 Group에 속한 문서들과의 Distance를 계산해, M은 그 중 가장 가까운 문서의 Group에 속하게 한다. 만약 M과 가장 가까운 문서와의 Distance가 MINSim 이하이면 M을 중심으로 새로운 Group을 형성한다. 이 경우에는 M과 아직 Group에 속하지 않은 문서들 간의 Distance를 구해, MINSim 이상인 문서들을 M과 같은 Group에 포함시킨다.

```
Site = {M0,M1,M2,.. Mt}
GroupV = {}  // Group에 속한 Matrix
GroupN = Site - GroupV;   // 아직 Group에 속하지 않은 Matrix

WHILE GroupN<>Empty DO
   M = SelectOneFrom(GroupN) // 검사할 문서. 아직 검사되지 않은 Site
   Boolean Group = False
```

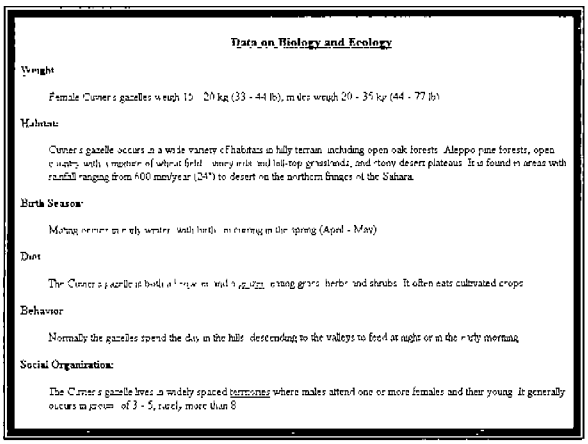
```
//M을 GroupV에 속한 Matrix들과 비교하면서, 가까운 Matrix가 있다면,
//그 Matrix가 포함되어 있는 그룹에 M을 포함시킨다.
WHILE (V=SelectOneFrom(GroupV)) <> NULL DO
   SIM = Dist(M,V)
   if(SIM > MINSim AND  SIM>D.MaxSIM)
      M.GroupID = V.GroupID
      M.MaxSIM = SIM
      Group = TRUE

ADD(GroupV,M)
//M이 Group에 속하지 않았다면
//M을 아직 그룹에 속하지 않은 문서들(N)을 비교하면서
//M과 가까운 문서 N들 새로운 그룹에 포함시킨다.
If Group==TRUE THEN
   GroupID = M.Number  // M.Number는 M의 일련번호
   WHILE (N=SelectOneFrom(GroupN)) <> NULL DO
      SIM = Dist(N,D)
      if(SIM > MINSim)
         N.GroupID = GroupID
         N.MaxSIM = SIM
         ADD(GroupV,N)
GroupN = Site - GroupV
```

(알고리즘 1) Matrix Clustering

4.3 Group Partitioning

Clustering을 하여 얻어진 Group 중에 어떤 Group은 문서들이 형식이 같아서 한 Group에 속했지만, 내용상으로는 다른 문서들이 혼합되어 있을 수 있다. 이런 Group은 다시 내용별로 나누어 주는 작업이 필요하다. 실험에서는 문서를 내용에 따라 전문 문서인지 아닌지를 분류하였다. 전문 문서 평가에는 간단한 조사를 이용하였다.

전문 문서 : 실험에서는 그림2와 같은 문서를 전문 문서라고 취급한다. 전문 문서를 그렇지 않은 문서와 구별하는 가장 간단한 방법은 <TAG> Properties of Animal </TAG>중에 하나라도 문서에 포함되어 있는지를 검사한다. <TAG> Properties of Animal </TAG>가 포함되어 있는 경우에는 문서가 전문 문서라는 표시를 해두었다.



Data on Biology and Ecology

Weight

Female Cuvier's gazelles weigh 15 - 20 kg (33 - 44 lb), males weigh 20 - 35 kg (44 - 77 lb)

Habitats

Cuvier's gazelle occurs in a wide variety of habitats in hilly terrain, including open oak forests, Aleppo pine forests, open country with a mixture of wheat field, stony wadi and hill-top grasslands, and stony desert plateaus. It is found in areas with rainfall ranging from 600 mm/year (24") to desert on the northern fringes of the Sahara.

Birth Season:

Mating occurs in early winter, with births occurring in the spring (April - May)

Diet

The Cuvier's gazelle is both a [illegible] eating grass, herbs and shrubs. It often eats cultivated crops.

Behavior

Normally the gazelles spend the day in the hills, descending to the valleys to feed at night or in the early morning.

Social Organization:

The Cuvier's gazelle lives in widely spaced territories where males attend one or more females and their young. It generally occurs in groups of 3 - 5, rarely more than 8.

(그림 2) 전문 문서의 예

한 Group안에 전문 문서가 없거나, 아니면 모두 전문 문서일 경우에는 Group의 처리가 쉽다. 그렇지만, 일부가

전문 문서라면 경우에 따라서는 Group을 나누어 주는 과정이 필요하다. Group k의 PR(Group내에 전문 문서들의 포함 비율)은 다음과 같이 계산한다

PR_k=Group k의 전문 문서 수/Group k의 문서 수

Group Partitioning은 Group k의 PR과 θ_1, θ_2의 값에 따라 표6과 같이 처리해 준다.

조건	Group Partitioning 여부
$PR_k < \theta_1$	나누지 않음
$\theta_1 <= PR_k < \theta_2$	전문 문서들로 구성된 Group k와 그렇지 않은 문서들인 Group k+1로 분할
$\theta_2 <= PR_k$	나누지 않음

(표 6) Group Partitioning 조건 ($0.0 < \theta_1 < \theta_2 < 1.0$)

4.4 Group Selection

Clustering을 하여 얻은 여러 Group중에 Group의 PR을 고려해 δ이상인 Group들에 속해 있는 문서들을 정보 에이전트의 Database에 저장한다. 단, 크기가 작은 Group은 제외한다.

조건	정보 에이전트의 Database에 저장
$PR < \delta$	저장하지 않음
$\delta <= PR$	저장함

(표 7) Group Selection 조건 ($0.0 < \delta < 1.0$)

5. 실험과 평가

5.1 문서 수집 결과

문서 수집의 효과를 평가하기 위해서, 다음의 세 Site를 임의로 선택해 방문해 보았다. 각 Site별로 문서 수집부터 Group Selection까지 한 결과는 표 8과 같다.

<선택된 Site들>
Site 가 : http://www.animalinfo.org
Site 나 : http://www.parks.tas.gov.au/
Site 다 : http://animaldiversity.ummz.umich.edu/

Site	가	나	다
전체 문서	471	1243	3571
전문 문서	210	58	995
선택된 문서	209	47	962
기존 RPN	37	13	65
선택된 문서-기존 RPN	172	34	897

(표 8) Site 방문결과
(Clustering : MINSim=0.8, θ_1=0.2, θ_2=0.8, δ=0.8)

표 8을 보면, Site '가'와 '다'의 경우에는 기존의 RPN보다 훨씬 많은 관련 문서를 찾아온 것을 알 수 있다. Site '가'는 기존의 관련 문서의 수보다 4.6배나 많은 새로운 관련 문서들을 찾아왔다. 특히 Site '다'는 기존의 관련 문서 수보다 13.8배나 많은 새로운 관련 문서들을 찾아 오는 효과를 보였다. 그렇지만, Site '나'와 같이 이미 대부분의 문서가 관련 문서로 정보 에이전트의 Database에 저장되어 있어서, 추가로 찾아진 새로운 관련 문서의 수가 적은 경우도 있다.

Site '가'는 전체 문서는 471개, 전문 문서는 210개를 가지고 있다. Site '다'는 전체 문서는 3571개, 전문 문서는 995개를 가지고 있다. 그렇지만, Site '나'는 전체 문서는 1243개, 전문 문서는 58개를 가지고 있다. 이를 보면, Site의 크기(전체 문서의 수)와 전문 문서의 수가 비례한다고 할 수 없다. 그렇지만, 방문한 Site에 있는 전문 문서 수가 상당히 크다는 것이 주목할 만하다.

Site의 전문 문서와 선택된 문서의 수에는 약간의 차이가 있는데, 이는 그 전문 문서를 포함한 Group의 크기가 작아서 제외되었기 때문이다.

5.2 외부로의 Link로의 Site Selection의 효과

외부로의 Link를 Site Selection의 한 요인인 LN으로 이용하는 것에 대한 효과를 평가하기 위해서, Site를 방문해서 얻은 외부로의 Link를 Site Table의 Site들과 비교해 겹치는 Site들과 새로운 Site들로 나누어 보았다.

Site	가	나	다
외부로의 Link	429	76	590
Site Table과 겹치는 Site	169 (39.4%)	75 (98.7%)	288 (48.8%)
새로운 Site	260	1	302

(표 9) Site를 방문해서 얻은 외부로의 Link의 평가

표 9에서 보는 바와 같이, 방문한 Site에서 얻어진 외부로의 Link가 상당수의 Site Table에 있는 관련 Site들을 Link하고 있음을 알 수 있다. 특히 Site '나'의 경우에는 76개의 외부로의 Link중에 75(98.7%)개가 겹친다. 이렇게 외부로의 Link가 Site Table에 있는 관련 Site들을 가리키는 것은 같은 영역의 Site들끼리 서로 밀접하게 연결되어 있다는 것을 의미한다. 그러므로 외부로부터 많이 Link된 Site일수록 그 영역에서 상당히 좋은 Site라고 여겨질 수 있다. 따라서, 외부로의 Link를 이용해 Site Selection을 하는 것은 타당하다고 볼 수 있다.

6. 결론

본 논문에서는 정보 에이전트가 Web에서 관련 문서를 효과적으로 찾기 위한 적극적 관련 문서 수집 기법에 대하여 설명하였다. 새로 제안된 기법에서는 Link와 Clustering을 이용한다. Link 중 내부로의 Link를 이용하여 Site의 문서들을 모아 오고, 외부로의 Link로는 Site Selection에 사용하거나 새로운 관련 Site들을 찾

아 낸다. 문서들은 기존처럼 개별적으로 평가•수집하는 것이 아니라, 문서 형식으로 Clustering하여 Group으로 처리한다. 실험은 실제 세 개의 Web Site를 방문해 보았다. 그 결과, 새로운 적극적 문서 수집 기법으로 숨겨져 있던 관련 문서들을 수집할 수 있었으며, 새로운 관련 Site들도 수집할 수 있었다.

본 연구는 뇌과학 연구 지원을 받아 진행되었습니다.

참고 문헌

[1] 이용현: 정보통신망에서 지능형 정보 에이전트와 특정 영역에서의 구현 (홍익대학교 박사학위논문) (1999)

[2] 김상모: 웹에서 특정영역 정보에이전트의 성능향상에 관한 연구. (홍익대학교 석사학위논문) (2000)

[3] Sergey Brin, Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. (1998)

[4] 서연규, 김경중, 정윤경, 조성배 : 웹사이트의 구조 분석을 위한 소프트웨어 에이전트 (2000)

[5] Jerome Moore, Eui-Hong (Sam) Han, Daniel Boley, Maria Gini, Robert Gross, Kyle Hastings, George Karypis, Vipin Kumar, and Bamshad Mobasher : Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering. (1997)