WordNet과 HTML 태그를 활용한 특정영역 정보의 웹 문서 분류

조은휘 , 변영태
홍익대학교 가상에이전트 연구실

Web Document Clustering for Specific Subject Information Using WordNet and HTML Tags

Eun-Hwi Cho , Young-Tae Byun
Virtual Agent Lab. , Hongik university

요 약

웹 상의 많은 정보들 속에서 사용자가 원하는 정보를 찾아내는 일은 쉽지 않다. 사용자가 의도하는 양질의 정보 제공을 위해 특정 영역과 관련한 정보 제공 시스템이 개발되고 있다. 이전 시스템은 특정 영역 관련 지식베이스를 토대로 하여 웹 문서를 수집해 놓고, 사용자에게 정보를 제공한다. 본 논문에서는 전문 사이트 내에 문서 간의 유사성을 토대로 하여 동물 영역에 대한 효과적인 문서 클러스타링(clustering)에 관해 실험하였다. 기존의 방법에서는 문서의 분류나 질의어와 관련한 문서 선택이나 순위 결정이 주로 텀(term)을 바탕으로 하고 있다. 본 논문에서는 각 문서 내의 텀 뿐만 아니라 HTML 태그(tag), 지식베이스에 WordNet의 계층구조를 적용한 data를 활용하고, SVD(Singular Value Decomposition)를 사용하여 문서 간의 관계를 밝혀내어 문서 분류 및 수집에 이용하였다. 특정 영역의 전문 문서를 많이 제공하는 사이트에 적용하여 좋은 결과를 볼 수 있었다.

1. 서론

점점 더 많은 양의 정보를 제공하고 있는 웹 상에서 보다 나은 양질의 정보를 검색하기 위한 정보 에이전트 시스템은 특정 영역에 관련된 웹 문서들을 수집해 놓았다가 사용자가 요청한 질의에 적합한 문서들을 제공해 준다[1].

이러한 정보 에이전트 시스템에서의 서비스 향상을 위해서는 좋은 웹 문서를 많이 가지고 있어, 사용자의 요구에 보다 알맞은 문서를 제공해 줄 수 있어야 한다.

웹 문서의 수집에 있어서, 기존에는 사용자의 질의를 확장하고, 이를 바탕으로 하여 주로 텀(term)을 위주로 하여 문서들을 분류, 수집하였다.

본 연구에서는 동물 영역과 관련한 전문적인 사이트들을 대상으로 실험하였다.

이러한 사이트들의 예는 다음과 같고, 이들은 동물과 관련한 속성 정보 및 지역 정보 등의 많은 정보들을 포함하고 있다.

http://www.animalinfo.org
http://www.parks.tas.gov.au
http://animaldiversity.ummz.umich.edu

실험에서는 http://www.animalinfo.org [Animal Info] 의 모든 웹 사이트를 대상으로 텀과 함께 HTML 태그(tag), WordNet의 계층 정보를 포함한 data를 활용한다.

이런 요소들로 이루어진 문서 정보를 가지고 SVD(Singular Value Decomposition)를 이용하여 문서들 간의 상관관계를 파악하고, 이 결과에 맞게 K-means clustering 기법을 변형하여 문서를 분류하였다.

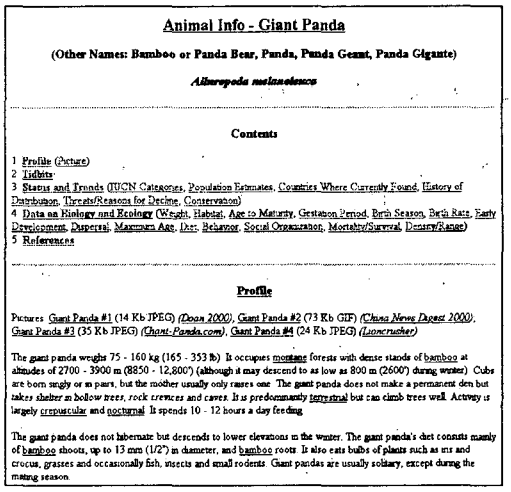
이 실험을 통해 유용한 웹 문서의 그룹을 찾아내는 데에 그 의의가 있다.

본 연구는 뇌과학 연구 사업의 지원으로 진행 되었음.

2. 문서 정보의 수집

동물 영역에 웹 문서 수집을 위해 동물에 대한 전문적인 정보를 제공하고 있는 대규모의 사이트 중 하나인 http://www.animalinfo.org : [Animal Info] 를 대상으로 실험하였다.

이 사이트는 전체 470개의 문서로 이루어져 있으며, 이들 중 동물의 정보를 설명하고 있는 <그림 1>과 같은 문서를 전문 문서로 보기로 한다.



Animal Info - Giant Panda

(Other Names: Bamboo or Panda Bear, Panda, Panda Geant, Panda Gigante)

Ailuropoda melanoleuca

Contents

1 Profile (Picture)
2 Tidbits
3 Status and Trends (IUCN Categories, Population Estimates, Countries Where Currently Found, History of Distribution, Threats/Reasons for Decline, Conservation)
4 Data on Biology and Ecology (Weight, Habitat, Age to Maturity, Gestation Period, Birth Season, Birth Rate, Early Development, Dispersal, Maximum Age, Diet, Behavior, Social Organization, Mortality/Survival, Density/Range)
5 References

Profile

Pictures: Giant Panda #1 (14 Kb JPEG) *(Doan 2000)*, Giant Panda #2 (73 Kb GIF) *(China News Digest 2000)*, Giant Panda #3 (35 Kb JPEG) *(Chant-Panda.com)*, Giant Panda #4 (24 Kb JPEG) *(Luoctrucker)*

The giant panda weighs 75 - 160 kg (165 - 353 lb). It occupies montane forests with dense stands of bamboo at altitudes of 2700 - 3900 m (8850 - 12,800') (although it may descend to as low as 800 m (2600') during winter). Cubs are born singly or as pairs, but the mother usually only raises one. The giant panda does not make a permanent den but takes shelter in hollow trees, rock crevices and caves. It is predominantly terrestrial but can climb trees well. Activity is largely crepuscular and nocturnal. It spends 10 - 12 hours a day feeding.

The giant panda does not hibernate but descends to lower elevations in the winter. The giant panda's diet consists mainly of bamboo shoots, up to 13 mm (1/2") in diameter, and bamboo roots. It also eats bulbs of plants such as iris and crocus, grasses and occasionally fish, insects and small rodents. Giant pandas are usually solitary, except during the mating season.

<그림 1>. 동물 정보를 가진 전문 문서

Animal Info 사이트 내에 이러한 동물 정보를 가진 전문 문서는 210개가 존재한다.

이 사이트 내의 문서 분류를 위해 텀, HTML 태그, WordNet의 계층 정보 data를 가지고 SVD를 사용하기 위해 Matrix를 생성하였다.

Matrix는 위 3가지 성분에 대한 출현 빈도수를 가지고 생성된다. 각 Matrix 생성에 대한 자세한 사항은 아래와 같다.

2.1 HTML 태그

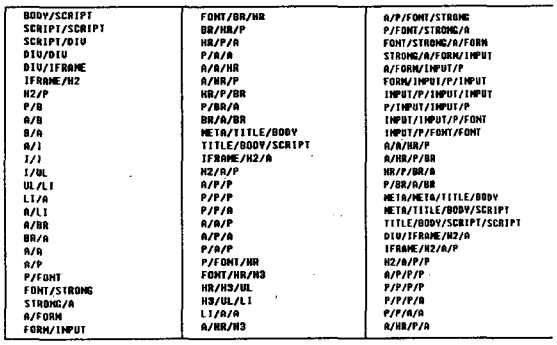
문서 내의 모든 태그들을 고려하여 Matrix를 생성한다. 각각의 태그 빈도에 관한 정보는 고려하지 않고, 연속된 태그 패턴에 주목하였다.

즉, 웹을 구성하고 있는 태그들에는 일정한 패턴이 있고 이것들은 웹 문서의 성격을 나타낼 수 있으므로 이점을 고려하여 태그의 sequence를 2개, 3개, 4개로 하여 그 빈도수를 구하였다.

<그림 2>는 이러한 2-gram, 3-gram, 4-gram Matrix의 정보를 보여주고 있다.

2.2 텀

분류하고자 하는 문서 집합 내의 모든 문서를



2-gram	3-gram	4-gram
BODY/SCRIPT	FONT/BR/HR	A/P/FONT/STRONG
SCRIPT/SCRIPT	BR/HR/P	P/FONT/STRONG/A
SCRIPT/DIV	HR/P/A	FONT/STRONG/A/FORM
DIV/DIV	P/A/A	STRONG/A/FORM/INPUT
DIV/IFRAME	A/A/HR	A/FORM/INPUT/P
IFRAME/H2	A/HR/P	FORM/INPUT/P/INPUT
H2/P	HR/P/BR	INPUT/P/INPUT/INPUT
P/B	P/BR/A	P/INPUT/INPUT/P
A/B	BR/A/BR	INPUT/INPUT/P/FONT
B/A	META/TITLE/BODY	INPUT/P/FONT/FONT
A/I	TITLE/BODY/SCRIPT	A/A/HR/P
I/I	IFRAME/H2/A	A/HR/P/BR
I/UL	H2/A/P	HR/P/BR/A
UL/LI	A/P/P	P/BR/A/BR
LI/A	P/P/P	META/META/TITLE/BODY
A/LI	P/P/A	META/TITLE/BODY/SCRIPT
A/BR	A/A/P	TITLE/BODY/SCRIPT/SCRIPT
BR/A	A/P/A	DIV/IFRAME/H2/A
A/A	P/A/P	IFRAME/H2/A/P
A/P	P/FONT/HR	H2/A/P/P
P/FONT	FONT/HR/H3	A/P/P/P
FONT/STRONG	HR/H3/UL	P/P/P/P
STRONG/A	H3/UL/LI	P/P/P/A
A/FORM	L1/A/A	P/P/A/A
FORM/INPUT	A/HR/H3	A/H3/P/A

<그림 2>. 2,3,4-gram Matrix Info.

고려하여 Matrix를 생성한다.

웹 문서 상의 단어들은 일반 문서와는 달리 사이트의 주소 또는 e-mail 등의 특정 정보들을 포함하고 있고, 다양한 사용자들이 정보를 제공하고 있으므로 많은 오탈자가 발견된다.

그러므로, 이를 고려하여 알파벳(alphabet)으로만 이루어진 텀을 고려하였고, 53570개의 단어 Dictionary와 Stoplist를 통하여 오탈자 제거 및 유용한 텀을 선택하였다.

웹 문서들 사이에서 텀에 빈도수를 고려하기 위해서 Stemming 작업도 함께 하였다.

$$X = U_0 S_0 V_0^T$$

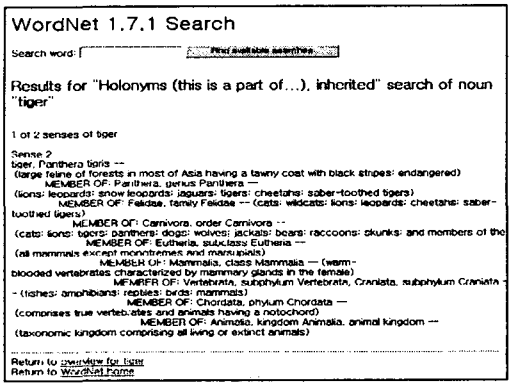
U_0 : Left Singular Vectors
S_0 : Singular Values - Diagonal Matrix
V_0 : Right Singular Vectors

2.3 워드넷(WordNet)

WordNet은 Ontology의 일종으로 인간의 어휘 지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스톤 대학 인지과학 연구실이 구축해 온 언어어휘 데이터베이스이다[4].

WordNet을 활용한 계층 정보 data는 문서 속

의 단어들 중에서 동물 영역과 관련한 단어들의 경우에 그 단어들을 "Web WordNet 1.7.1"을 사용하여, 상위 animal 까지의 사이에 있는 단어들을 모두 가져와서 구성한다.



<그림 3>. WordNet: 'tiger' 검색 화면

3. 문서 분류

문서가 가지고 있는 태그, 텀, WordNet data를 가지고 생성한 Matrix를 가지고, 문서를 분류해 내기 위해 SVD를 사용하여 문서 간의 상관관계를 구해내고, 이 관계를 가지고 Clustering 하여 문서를 분류하였다.

3.1 SVD (Singular Value Decomposition)

SVD는 우리가 가지고 있는 Matrix를 다음과 같은 성분으로 나누어서 생각한다[2][3].

이것은 Matrix를 구성하는 값들 중 그 중요도에 따라 S Matrix의 value에서 일부를 취해 새로운 Matrix X′ 을 생성해 낼 수 있다[2].

SVD를 적용한 X′ Matrix를 사용하여 문서 간의 유사성을 파악하기 위해서는 아래의 식을 사용한다[3].

$$X^TX = (USV^T)^TUSV^T = VS^TU^TUSV^T = (SV^T)^T(SV^T)$$

위 식을 사용하면 각각의 문서들 간의 관계를 나타내는 (doc# * doc#) 크기의 문서를 얻을 수 있다.

3.2 K-means Clustering

본 연구의 실험에서는 기본적으로 K-means 알고리즘에 바탕을 두고 수정한 clustering 방법을 사용한다.

즉, 문서 간에 상관관계가 가장 적은 점들을 선택하여 clustering의 seed로 고정하고, 이 seed를 중심으로 문서들을 분류해 나간다.

실험에서는 SVD를 사용한 문서들 간의 관계가 값으로 나타나게 되고 이 값이 클수록 강한 상관관계를 가지게 되는데, 이때 자기자신과의 관계가 다른 문서와의 관계보다 작은 값을 가지게 되는 경우가 존재하므로, 하나의 seed가 다른 seed에게 종속되어질 수 있다. 그러므로 seed의 개수와 group의 개수가 일치하지는 않는다. (실험에서의 시작 seed 개수는 100으로 두었다.)

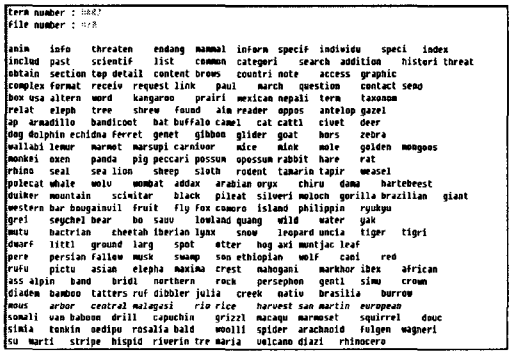
4. 실험결과 및 평가

(1) Tag
: (2907 * 470) // 264+ 825+ 1818

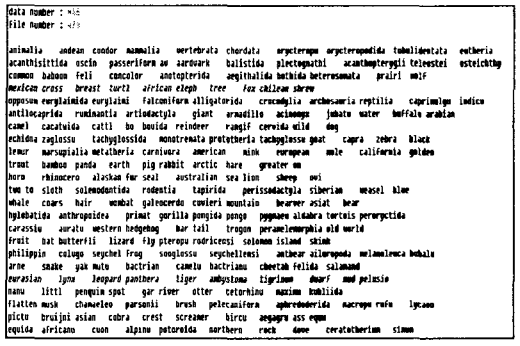
$$X' = USV^T$$

(2) Term
: (4487 * 470)
(3) WordNet
: (846 * 470)
(4) Tag + Term
: (7394 * 470) // 2907+ 4487
(5) Tag + WordNet
: (3753 * 470) // 2907+ 846
(6) Term + WordNet
: (5333 * 470) // 4487 + 846

참고로 <그림 4>는 Term에 관한 정보를, <그림 5>는 WordNet에 관한 정보를 보여준다.



<그림 4>. Term Info.



<그림 5>. WordNet data Info.

다음의 <표 2>는 (1)~(6)까지 6가지 경우에 대한 문서 분류 결과를 보여주고 있다. 이는 Clustering의 seed를 100으로 두었을 때, 생성되는 그룹(group)에서,

(전문문서의 수 / 그룹의 문서 수)

를 보이고, 이들 중 전문문서를 포함하고 있는 경우를 음영으로 나타내었다.

Group	(1) Tag	(2) Term	(3) WN
1	74 / 260	3 / 9	6 / 33
2	0 / 54	11 / 71	0 / 1
3	0 / 17	2 / 47	4 / 5
4	136 / 137	0 / 4	200 / 431
5	0 / 2	6 / 7	
6		0 / 4	
7		142 / 153	
8		0 / 27	
9		31 / 31	
10		0 / 2	
11		15 / 115	

<표 2-1>. 결과 Table 1.

Group	(4)Tag+Term	(5)Tag+WN	(6)Term+WN
1	0 / 54	152 / 391	2 / 16
2	0 / 2	0 / 15	7 / 8
3	13 / 131	58 / 58	10 / 57
4	0 / 4	0 / 2	0 / 2
5	0 / 13	0 / 4	6 / 52
6	32 / 32		20 / 50
7	0 / 1		21 / 21
8	7 / 47		34 / 34
9	62 / 62		89 / 90
10	0 / 10		0 / 27
11	74 / 77		6 / 6
12	5 / 5		15 / 107
13	0 / 1		
14	0 / 27		
15	0 / 4		

<표 2-2>. 결과 Table 2.

4.2 평가

문서의 그룹이 잘 분류되기 위해서는 생성되는 그룹에서 전문문서를 포함하는 경우가 명확히 구분되어야 하고, 그런 그룹이 되도록 적어야 한다.

<표 2-1>의 결과 Table을 보면, Tag의 경우가 Clustering이 가장 잘 되는 것을 볼 수 있다. 이와 달리 Term의 경우는 group이 많이 분산되는 것을 살펴볼 수 있다.

<표 2-2>는 각각의 요소들을 함께 적용하였을

때의 결과이다. 여기에서는 Tag와 WN(WordNet)을 함께 적용하였을 때, 좋은 결과를 볼 수 있었다.

이 결과에서 보면, HTML 태그의 정보가 매우 유용하다는 것을 알 수 있다.

5. 결론 및 향후 과제

본 연구에서는 동물 영역과 관련한 하나의 대규모 사이트에서는 HTML 태그의 정보가 유용하다는 사실을 볼 수 있다.

그러나, 이 경우의 사이트는 모든 문서가 동물과 관련한 정보를 가지고 있었으므로 팀이나 WordNet의 data가 제대로 활용되지 못했을 수 있다.

하지만, 이 실험은 기존의 경우와는 다리 어떤 부분에 특화하여 구현된 것이 아니라 웹 문서 상의 가능한 한 모든 자료들을 고려하고 있으므로 보다 많은 웹 문서의 분류에 적용 가능할 것이다.

그러므로, 다음에는 이러한 점을 고려하여 다른 주제와 관련한 사이트나, 여러 주제를 포괄하고 있는 사이트에서의 실험을 하여 각각의 모든 정보를 고려한 문서의 분류 작업이 가능할 것이다.

또한 SVD나 clustering 시에 사용되는 dimension과 seed의 개수를 변화시켜 생각해 볼 수도 있을 것이다.

참고 문헌

[1] 이용현, "정보통신망에서 지능형 정보 에이전트와 특정 영역에서의 구현", 홍익대학교 박사학위 논문, 1999

[2] Scott Derrwester, Susan T. Dumals, George W. Furnas, Thomas K. Landauer, Richard Harshman, "Indexing by Latent Semantic Analysis", Jornal of the American Society of Information Science, 1990

[3] 김준태, "추천 시스템에서의 차원 축소 효과", 지능형 에이전트 워크샵, 2002

[4] http://www.cogsci.princeton.edu/~wn/

[5] http://math.nist.gov/javanumerics/jama/

[6] 유민홍, "워드넷을 이용한 통계 기반의 영어 복합명사 분석", 연세대학교 대학원, 2000