전문 웹 사이트에서의 통계적 기법과 태그 정보를 이용한 문서 분류

조은휘, 변영태
홍익대학교 컴퓨터공학과

Web Document Clustering Using Statistical Techniques & Tag Information on the Specific-Domain Web site

Eun-Hwi Cho , Young-Tae Byun
Dept. of Computer Science, Hongik University

요 약

특정 영역에 대해 사용자에게 관련 정보를 제공하는 서비스를 위해 정보 에이전트를 개발하고 있다. 이 시스템은 웹 상에서 문서를 수집해 오는데 특정 영역과 관련한 지식베이스를 토대로 하고 있는데, 이들 중 몇몇 전문 사이트 내의 정보가 많이 포함되어 있음을 볼 수 있다. 그러므로 전문 사이트 내의 관련 문서 수집은 중요한 의의가 있다. 본 논문에서는 이들 전문 사이트 내의 전문 문서 수집을 위해 문서 간의 유사성을 토대로 클러스터링 한다. 즉, 문서 내의 텀(term)과 HTML 태그(tag), 지식베이스의 WordNet 계층구조를 data로 하고 SVD(Singular Value Decomposition)을 사용하여 문서 간의 관계를 밝혀내었다.

1. 서론

특정 영역 정보 에이전트는 사용자가 원하는 정보를 담고 있는 관련 웹 문서를 제공해 주는 서비스 수행한다. 따라서 특정 영역 관련 웹 문서를 많이 수집할 수 있어야 한다[1].

정보 에이전트의 Database를 분석한 결과 관련 문서들이 있는 사이트에서 방문하지 않은 관련 문서들이 더 많이 있다는 것이 발견되었고, 기존 관련 문서가 있는 사이트에서 숨겨져 있는 문서를 찾아내는 방법이 필요하게 되었다[2].

이번 연구는 동물 영역과 관련한 사이트에서 관련 문서들을 찾아내기 위해 기본적으로 문서 내의 텀(term)과 함께 HTML 태그(tag), 지식베이스의 WordNet 계층구조를 data로 하여 통계적 기법인 SVD(Singular Value Decomposition)을 사용한다.

실험에서는 영문 사이트인 http://www.animalinfo.org [Animal Info] 와 한글 사이트 http://www.nature21.com [네이처21] 을 그 대상으로 하였다.

2. 실험 방법

실험은 각 사이트('Animal Info', '네이처21')를 구성하고 있는 웹 문서를 대상으로 각 문서의 텀과 함께 HTML 태그, WordNet의 계층 정보를 data로 한다. 이 세 종류의 data를 가지고 SVD(Singular Value Decomposition)을 이용하여 문서들 간의 상관관계를 파악하고, SVD를 통한 문서간의 관계 값을 가지고 K-means Clustering 기법을 변형하여 문서를 분류해 내었다.

2-1. 문서 정보

동물 영역과 관련된 영문 사이트 [Animal Info]의 470개 문서와 한글 사이트 [네이처21]의

본 연구는 뇌과학 연구 사업의 지원으로 진행 되었음.

150개 문서에서 텀, HTML 태그, WordNet의 계층 정보를 data로 SVD를 실행하기 위한 Matrix를 생성한다.

2-1-1. 텀

분류하고자 하는 사이트 내의 모든 문서를 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 영문 사이트에서는 알파벳(alphabet)만으로 이루어진 텀을 선택, Dictionary 및 Stoplist를 통해 오탈자 제거 및 유효한 텀을 선택하였고, 한글 사이트는 고려대의 자연어 처리 모듈인 '한국어 형태소 분석 및 품사 부착 시스템'[4]을 이용하여 일반 명사와 고유 명사만을 선택하였다.

2-1-2. HTML 태그

문서 내의 모든 태그들을 고려하여 Matrix를 생성한다. 태그 각각에 대한 빈도는 고려하지 않고, 연속된 태그 패턴 정보만을 고려하였다.

즉, 웹 문서를 구성하고 있는 태그들에는 일정 패턴이 있고 이것이 웹 문서의 성격을 나타낼 수 있다는 점에 주목하여 태그 sequence를 2개, 3개, 4개로 하여 그 빈도수를 구하였다.

<그림1>은 2, 3, 4개의 HTML 태그 데이터 정보를 보여주고 있다.

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	LI/A/A	P/P/A/A
FORM/INPUT	A/HR/H3	A/HR/P/A

<그림1> 2,3,4-gram HTML 태그 정보

2-1-3. WordNet 계층 정보

WordNet은 Ontology의 일종으로 인간과 어휘 지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스톤 대학 인지과학 연구실이 구축해 온 언어어휘 데이터베이스이다[3].

WordNet을 활용한 계층 정보 data는 문서 속의 단어들 중에서 동물 영역과 관련한 단어들의 경우에 그 단어들을 "Web WordNet 1.7.1"을 사용하여, 상위 animal 까지의 사이에 존재하는 단어들 모두를 가져와서 구성한다.

2-2. 문서 분류

문서가 가지고 있는 텀, 태그, WordNet의 계층 정보 data를 가지고 생성한 Matrix를 가지고, 문서를 분류해 내기 위해 SVD를 사용하여 문서 간의 상관관계를 구해내고, 이 관계를 가지고 Clustering하여 문서를 분류하였다.

2-2-1. SVD (Singular Value Decomposition)

SVD는 data Matrix를 다음과 같은 성분으로 나누어서 생각한다[5][6].

$$X = U_0 S_0 V_0^T$$

U_0 : Left Singular Vectors

S_0 : Singular Values - Diagonal Matrix

V_0 : Right Singular Vectors

이것은 Matrix를 구성하는 값들 중 그 중요도에 따라 S Matrix의 value에서 일부를 취해 새로운 Matrix X′ 을 생성해 낼 수 있다[5].

$$X' = USV^T$$

SVD를 적용한 X′ Matrix를 사용하여 문서 간의 유사성을 파악하기 위해서는 아래의 식을 사용한다[6].

이 식을 사용하면 각각의 문서들 간의 관계를 나타내는 (doc# × doc#) 크기의 Matrix를 얻을 수 있다.

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

2-2-2. Clustering

이 실험에서는 기본적으로 K-means 알고리즘에 바탕을 두고 수정한 Clustering 방법을 사용한다.

문서 간의 상관관계가 가장 적은 점들을 선택하여 Clustering의 seed로 고정하고, 이 seed를 중심으로 문서들을 분류해 나간다.

SVD는 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 일치하지 않을 수 있다.

이 알고리즘을 정리하면 다음과 같다.

function Kmeans(int k)
{
initialization
select seed group $S = \{ S_1, S_2, \ldots, S_k \}$ $(k < n)$
from document $d_1, d_2, \ldots, d_n$
$\leftarrow$ *min* ($rel(d_i, d_j)$)
clustering
for each input document dj
where ($j \ni \{1, 2, \ldots, n\}$)
do
given d_j, *found* $S_i \leftarrow min \mid d_j - S_i \mid$
if $d_j \notin S$, $d_j \in G_i$
$d_j \in S$, $G_i \leftarrow G_i \cup G_j$
$S \leftarrow S - S_j$

3. 실험결과 및 평가

[Animal Info] 사이트의 470개 문서와 [네이처21]의 150개 문서에 대한 실험 data 개수는 아래의 <표1>과 같다.

	[Animal Info]	[네이처21]
Tag	2907	1237
Term	4487	2054
WordNet	846	
Tag + Term	7394	3291
Tag + WordNet	3753	
Term + WordNet	5333	

<표1> 각 사이트의 실험 data 개수

([네이처21]의 경우 WordNet의 한글 정보가 구축되지 않아, Tag, Term, Tag+Term의 경우만을 고려하였음.)

3-1. 실험결과

다음 <표2-1>, <표2-2>는 [Animal Info]의 문서 분류 결과이고, <표3>은 [네이처21]에 대한 문서 분류 결과를 보여주고 있다. [Animal Info]의 경우 SVD의 dimension을 10으로, Clustering의 seed를 100으로 두었고, [네이처21]의 경우는 dimension = 10, seed = 50으로 정하고 실험하였다.

표는

(전문문서의 수 / 그룹의 문서 수)

를 보이고, 이들 중 전문 문서를 포함하고 있는 경우를 음영으로 나타내었다.

Group	Tag	Term	WordNet
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	

<표2-1> [Animal Info] 결과 Table(1)

Group	Tag + Term	Tag + WN	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		

<표2-2> [Animal Info] 결과 Table(2)

Group	Tag	Term	Tag + Term
1	0 / 1	1 / 4	0 / 1
2	5 / 5	20 / 23	5 / 5
3	1 / 1	0 / 1	1 / 1
4	44 / 143	0 / 18	44 / 143
5		19 / 49	
6		3 / 5	
7		3 / 4	
8		3 / 8	
9		1 / 38	

<표3> [네이처21] 결과 Table

3-2. 평가

문서의 그룹이 잘 분류되기 위해서는 생성되는 그룹에서 전문문서를 포함하는 경우가 명확히 구분되어야 한다.

이를 측정하기 위해 다음 수식과 같은 엔트로피(Entropy) 값을 사용한다.

- Entropy 1.

$$E_1 = \frac{1}{m}\sum_m \sigma_m \qquad (m : \#\ of\ group)$$

$$\sigma_N = -\sum \frac{freq(C_i,N)}{N} \times \log_2 \left(\frac{freq(C_i,N)}{N}\right)$$

- Entropy 2.

$$E_2 = \sum_m \left(\sigma_m \times \frac{k}{T} \right)$$

$(k : \#\ of\ group\ ,\ T : \#\ of\ total\ document)$

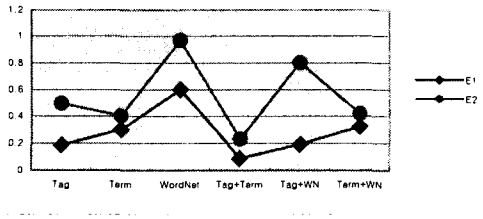
엔트로피의 값은 그 값이 작을수록 그룹핑이 더 잘되었음을 나타낸다.

Entropy 1.은 그룹의 개수로 나누어 엔트로피를 계산한 것이고, Entropy 2.는 그룹의 크기를 고려한 엔트로피 값이다.

<표4>와 <그림2>는 [Animal Info]에 대한 엔트로피 값이고 <표5>와 <그림3>은 [네이처21]에 대한 엔트로피의 값이다.

	Tag	Term	WN	Tag+Term	Tag+WN	Term+WN
E1	0.185	0.302	0.601	0.087	0.193	0.326
E2	0.495	0.404	0.969	0.230	0.802	0.419

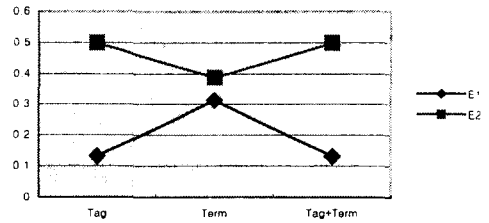
<표4> [Animal Info] 엔트로피



<그림2> [Animal Info] 엔트로피

	Tag	Term	Tag+Term
E1	0.131	0.312	0.131
E2	0.499	0.387	0.499

<표5> [네이처21] 엔트로피



<그림3> [네이처21] 엔트로피

[Animal Info]의 엔트로피는 'Tag+Term'의 경우가 가장 좋았고, 'Tag'의 경우에도 좋은 결과를 얻을 수 있었다. 그러나, 'WordNet'만을 사용하는 경우 가장 좋지 않았다.

[네이처21]의 경우 'Tag'의 영향으로 'Tag+Term'도 'Tag'만 있는 경우와 같은 결과가 발생하였으며 이 경우에 엔트로피1은 'Tag'의 경우가 더 나은 결과가 엔트로피2는 'Term'의 경우 더 좋은 결과를 얻을 수 있었다.

이는 [Animal Info]의 경우 문서의 분류에 있어 HTML 태그가 유용하다는 것을 보여주고 있다. 즉, 문서의 기술방법이 동물 정보를 포함하고 있는 경우 다른 문서들의 HTML 태그와는 다른 구조로 나누어져 있음을 말한다. 그리고 실제 사이트를 보더라도 이와 같은 양상을 띤다는 것을 살펴볼 수 있다.

그러나, [네이처21]의 경우에는 태그 정보를 적용하는 것이 좋은 방법이 되지 못하였다. 이것은 [네이처21]이라는 한 사이트 내의 문서들 대개가 비슷한 태그 형태를 지니고 있기 때문이라고 생각해 볼 수 있다. 그러나 이 경우 엔트로피1의 경우는 그 결과값이 조금 다르게 나타나고 있는데, 이것은 태그를 이용한 분류에서 하나의 그룹에 대개의 문서가 포함되어져, 텀의 나누어지는 그룹 개수가 상대적으로 커졌기 때문일 것이다.

이러한 결과는 인터넷 상의 모든 전문 문서들을 한가지 방법을 통해 분류하는 것은 어렵다는 것을 보여준다. 즉, 각 사이트의 기술 특성에 따라 태그의 정보가 텀보다 유용해질 수도 있고, 반대로 텀의 정보가 더 유용하게 쓰일 수도 있을 것이다.

3-3. 기타

SVD와 Clustering을 이용하는데 있어서, SVD에서는 dimension의 값이 하나의 가변 요인이 되고, Clustering을 사용하는 경우에는 seed의 개수가 하나의 가변 요인이 될 수 있다.

그렇다면, 각각의 경우에 그 값이 어떻게 변화하는지 [Animal Info]의 예를 통해 간단히 살펴보았다.

3-3-1. Dimension Factor 변화

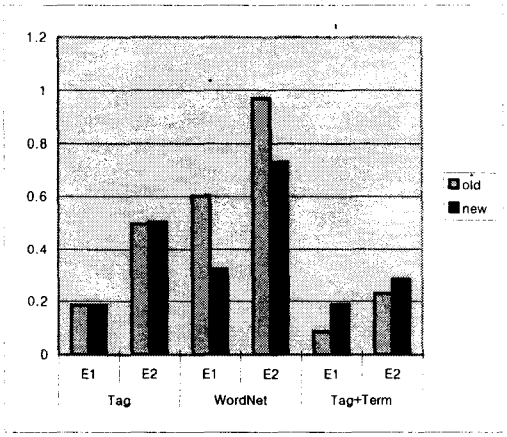
[Animal Info]의 Tag, WordNet, Tag+Term의 경우에 eigen vector의 변화 추이를 살펴보고 이 값이 크게 변화하지 않는 부분을 택하여 기존 10의 값에서 <표6>의 값과 같이 dimension을 조정하였고, 이후의 엔트로피 값의 변화는 <표7>과 같다.

	Dimension	
	Old	
Tag	10	13
WordNet	10	5
Tag+Term	10	12

<표6> [Animal Info] dimension factor의 변화

	Tag		WordNet		Tag+Term	
	E1	E2	E1	E2	E1	E2
Old	0.185	0.495	0.601	0.969	0.087	0.230
	0.186	0.502	0.323	0.729	0.188	0.284

<표7> [Animal Info] 엔트로피의 변화



<그림4> [Animal Info] 엔트로피의 변화

<표7> 엔트로피의 변화를 <그림4>에서 보다 쉽게 볼 수 있다.

이것을 보면 SVD의 dimension의 변화가 'WordNet'의 경우에는 엔트로피의 값을 작게 변화시키지만, 'Tag+Term'의 경우에는 반대로 증가시키는 것을 볼 수 있다.

SVD의 dimension은 많은 데이터들을 판단하는데 있어 고려되어지는 factor의 개수를 한정시키는 역할을 하여주는데, 'WordNet'의 경우는 이 값을 줄여서 더 좋은 결과를 얻을 수 있었고, 'Tag+Term'의 경우에는 그 값을 늘리는 경우를 택하였는데, 이 경우에는 엔트로피가 증가하였다.

즉, dimension의 변화는 엔트로피의 값을 변화시키는 요인으로 문서 분류에 영향을 끼친다는

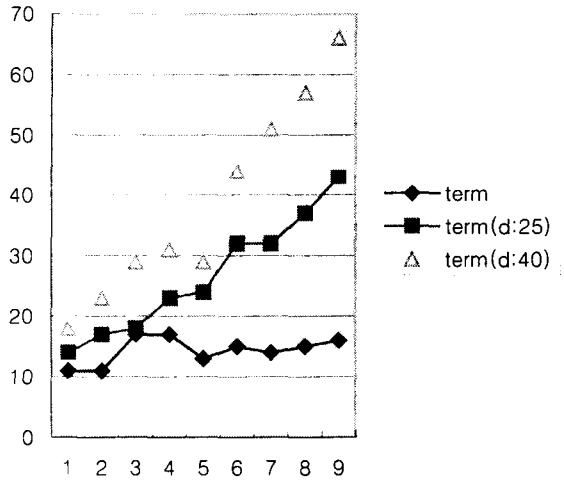
것을 볼 수 있지만, 그 값을 결정하는 문제가 보여진다.

3-3-2. Clustering의 Seed 변화

[Animal Info]사이트에서 Clustering 과정에서 seed의 개수를 50, 100, ..., 450까지 50씩 변화를 주었을 때, Term과 Term의 dimension이 25, 40일 때의 그룹 개수를 살펴보았다. (<표8>, <그림5>)

Seed#	Term	Term(d:25)	Term(d:40)
50	11	14	18
100	11	17	23
150	17	18	29
200	17	23	31
250	13	24	29
300	15	32	44
350	14	32	51
400	15	37	57
450	16	43	66

<표8> [Animal Info] seed에 따른 그룹 개수 변화



<그림5> [Animal Info] seed에 따른 그룹 개수 변화

그룹의 개수는 seed 개수 증가에 따라 전체적으로는 비례하는 양상을 띤다. 그러나 계속적으로 증가하는 것이 아니라 중간에 그룹의 개수가 다시 작아지거나 정체되는 구간이 발생한다.

하지만 전체적으로 볼 때, Clustering의 Seed의 값은 전체 테스트하는 data의 값의 중간치 보다 작은 값을 잡아주면 적절하다고 보여진다.

4. 결론 및 향후과제

이제까지 사이트 내의 관련 문서를 추출해 내기 위해 웹 사이트 내의 문서 분류 방법에 대해 살펴보았다.

우선, 문서 분류를 위해 사용되는 data의 값으로 텀 및 HTML 태그와 WordNet의 계층구조를 함께 사용하였다. 이 같은 data들을 가지고, 통계적 기법의 SVD를 사용하였으며 여기서 얻은 문서간의 상관관계에 대한 수치를 이용하여 Clustering 하였다. 그리고 이를 평가하기 위해 엔트로피를 구하여 보았다.

결과적으로 본다면 웹 사이트 내의 문서 분류는 각각의 사이트의 특성에 맞게 알맞은 방법을 도입해야 하고, 각 단계를 거치는 경우에 사용되는 다양한 요인들에 대한 값에 따라 그 결과치에 영향을 미친다는 것을 볼 수 있었다.

그러나, 다른 무엇보다도 하나의 웹 사이트를 분류하는데 있어서 그 사이트가 어떤 방법으로 기술되어 있는지에 대해 올바르게 이해하고 접근하는 것이 바람직하다고 보여진다. 즉, 기술하는 방법이 단순히 기술되는 내용에 따른 단어 그룹, 텀에 의하여 이루어지는지 또는 문서의 보여지는 형태, 태그의 변화에 따라 이루어지는지에 대한 결정을 통해 문서 분류의 기법을 적절하게 선택하는 것이 필요하다.

5. 참고문헌

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

[2] 김원우, "Link와 Clustering을 이용한 적극적 문서 수집 기법", 2001

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

[4] http://nlp.korea.ac.kr

[5] 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

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