Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 2, April 2013, pp. 126-132
http://dx.doi.org/10.5391/JKIIS.2013.23.2.126

웹페이지의 의학용어 출현 빈도와 하이퍼링크에 기반한 웹사이트 분류

Website Classification based on Occurrence Frequency of Medical Terms and Hyperlinks in Webpage

이인근* · 김화선** · 조훈*†
In Keun Lee · Hwa Sun Kim · Hune Cho†

*경북대학교 의료정보학과, **대구한의대학교 IT의료산업학과
† Department of Medical Informatics, Kyungpook National University
Department of Medical Information Technology, Daegu Haany University

요 약

본 논문은 웹페이지에 포함된 의학용어의 출현 빈도와 웹페이지 간의 하이퍼링크로 이루어진 웹사이트의 구조에 기반하여 인터넷 웹사이트를 분류하는 방법을 제안한다. 제안하는 방법에서는 (1)웹페이지에 포함된 전체 용어에서의 의학용어 출현 빈도와 (2)웹페이지에 포함된 중복을 제거한 용어에서의 의학용어 출현 빈도를 인자로 하여 웹페이지의 의학분야 적합도를 측정한다. 그리고 (3)홈페이지로부터 특정 웹페이지에 접근하기 위해 거쳐야 하는 하이퍼링크의 개수를 이용한 전체 웹페이지의 적합도 연산을 통해 웹사이트의 의학분야 적합도를 측정한다. 인터넷 포털 사이트의 디렉토리 검색 서비스에 등록된 80 개의 의학분야 웹사이트와 127 개의 비 의학분야 웹사이트를 대상으로 제안한 방법에 기반하여 웹사이트 분류 실험을 수행하였고, 82.5 %의 분류 정확률을 확인하였다.

키워드 : 웹사이트 분류, 용어 출현 빈도, 웹사이트 구조, 적합도 척도

Abstract

This study proposed a method to classify internet websites based on occurrence frequency of medical terms in the webpages and website structure composed with webpages and hyperlinks. The classification was done by using the suitability measure defined by three factors: (1)occurrence frequency of medical terms in the whole terms involved in a webpage, (2)occurrence frequency of medical terms in de-duplicated terms involved in the webpage, and (3)the number of hyperlinks to reach to a specific webpage from homepage. We conducted an experiment to verify the proposed method with the 80 websites registered in directories related to medical field and 127 websites in nonmedical field directories, and the experiment result showed 82.5 % of accuracy of the classification.

Key Words : Website Classification, Term Frequency, Website Structure, Suitability Measure

1. 서 론

일반적인 인터넷 검색 사이트에서는 검색어(keyword) 기반의 웹페이지(webpage) 검색 서비스(이하 “키워드 검색”)를 통해 검색어를 포함하는 웹페이지만을 선별하여 검색자에게 제공하고 있다. 그러나 인터넷에서의 웹페이지 양이 증가함에 따라 키워드 검색을 통해 검색되는 웹페이지의 양 또한 늘어남으로써 키워드 검색 방법만으로는 효과적인 검색 서비스를 제공할 수 없게 되었다. 이러한 문제점으로 인해 Google과 같은 검색 시스템에서는 PageRank와 같은 웹페이지 순위 정렬 알고리즘을 이용하여 특정 검색어를 포함하는 웹페이지 중에서 사용자의 검색 의도에 가장 적합하다고 판단하는 웹페이지들을 상위에 배열하고 있다. 또한, 분야별로 분류해 둔 웹페이지를 대상으로 키워드 검색을 수행함으로써 검색 정확률을 높이는 방법도 제안되었다[1,2,3].

접수일자: 2012년 11월 30일
심사(수정)일자: 2013년 2월 4일
게재확정일자 : 2013년 2월 8일
† Corresponding author
감사의 글 : 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (No.2012-004829), 이 논문은 2012년도 경북대학교 학술연구비에 의하여 연구되었음

특히, 한국어 문서 및 웹페이지의 분류 연구에서는 문서에 포함된 단어의 출현 빈도를 이용한 학습 및 규칙 기반의 문서 분류 방법들이 연구되어 왔다[4-9]. 즉, 학습 기반의 문서 분류 연구에는 벡터 공간 모델을 이용한 문서 분류[4], 신경망을 이용한 문서 분류[5], SVM (Support Vector Machine) 알고리즘을 이용한 문서 분류[6], 그래프 기반의 순위화 알고리즘(TextRank)을 이용한 문서 분류[7], Naïve Bayes 방법론을 이용한 개인정보 분류[8] 등이 있으며, 이들 연구에서는 90 % 내외의 문서 분류 정확률을 보인다. 또한, 한국어 어휘사전(U-WIN)에 기반하여 단어의 출현 빈도(Term Frequency)와 단어 사이의 관계 정도(Relation Value)를 이용한 문서 분류 연구[9]와 같은 규칙 기반의 문서 분류에서는 평균 84.4 %의 문서 분류 정확률을 보였다.

키워드 검색을 통한 웹페이지 접근 방법과 달리, 웹사이트(website)를 통한 정보접근은 검색자가 검색 분야와 관련된 웹페이지들을 그룹별로 확인할 수 있고, 웹사이트에서 최근 수정된 웹페이지 정보를 바로 확인할 수 있으며, 관련된 주제의 정보들을 포함한 웹사이트들을 상호 비교할 수 있는 장점이 있다. 웹사이트를 통한 정보접근의 이러한 장점으로 인해, Yahoo나 Open Directory Project와 같은 디렉토리(directory) 기반의 검색 서비스에서는 유사한 웹페이지들을 포함하는 웹사이트들을 그들의 성격에 따라 미리 분류해 둠으로써, 검색자가 정련된 웹사이트 정보를 얻을 수 있도록 하였다. 그러나 디렉토리 기반의 검색 서비스에서는 웹사이트 제공자가 특정 디렉토리를 선택하여 직접 웹사이트를 등록해야하기 때문에, 검색자는 디렉토리에 등록되지 않은 웹사이트를 찾을 수 없는 단점이 있다. 따라서 웹페이지를 분석하여 자동으로 분류하는 방법뿐만 아니라, 웹페이지를 포함하는 웹사이트를 자동으로 분류하는 다양한 방법이 연구되고 있다[10-13]. 이들은 Naïve Bayes 분류기를 이용한 웹사이트 분류[10], Hidden Markov Tree Model 기반의 웹사이트 분류[11], k-NN(k-nearest neighbor) 알고리즘을 이용한 웹사이트 분류[12]와 같이 사전 학습을 통한 웹사이트 분류 방법에 관한 연구로써, 최대 87 %의 웹사이트 분류 정확률을 보인다. 그러나 학습에 기반한 웹사이트 분류는 학습 데이터에 따라서 웹사이트의 분류 정확률이 결정되며, 학습 과정을 위한 학습 데이터의 선정 및 분석에 많은 노력과 시간이 소요되는 단점이 있다. 또한 하이퍼링크(hyperlink) 정보로 구성된 웹사이트의 구조에 기반하여 웹사이트를 분류하는 연구[13]에서는 최대 85 % 의 웹사이트 분류 정확률을 보이나, 웹페이지가 나타내는 정보의 의미를 고려하지 않고 있다.

본 연구에서는 사전에 구축된 분야 용어 사전과 웹사이트에 포함된 웹페이지들 간의 하이퍼링크 정보를 이용하는 규칙 기반의 웹사이트 분류 방법을 제안한다. 제안한 방법에서는 (1)웹페이지에 포함된 전체 용어에서 분야 용어의 출현 빈도와 (2)웹페이지에 포함된 중복을 제거한 용어에서 분야 용어의 출현 빈도를 인자로 하여 웹페이지의 분야 적합도를 계산한다. 그리고 (3)홈페이지[1]로부터 특정 웹페이지에 접근하기 위해 거쳐야 하는 하이퍼링크의 개수와 전체 웹페이지의 적합도 연산을 통해 웹사이트의 의학분야 최종 적합도를 결정한다. 제안한 방법의 검증을 위해 의학분야의 한국어 웹사이트 분류 실험을 수행하였다. 실험에서는 인터넷 포털 사이트의 디렉토리 검색서비스에 등록된 80 개의 의학분야 웹사이트에서 4022 개의 웹페이지, 그리고 127 개의 비 의학분야 웹사이트에서 7176 개의 웹페이지를 수집하여 웹사이트의 분야 적합도를 측정하였다. 실험결과, 82.5 %의 웹사이트 분류 정확률을 확인하였다.

2. 웹사이트 분야 적합도

본 연구에서는 "특정 분야에 관련된 웹페이지는 분야 용어를 많이 포함할 것이며, 특정 분야에 관련된 웹사이트에서는 분야 정보를 많이 포함한 웹페이지가 홈페이지로부터 접근하기 쉽도록 구성되어 있다"는 것을 전제로 하여 웹페이지 및 웹사이트의 분야 적합도를 측정한다. 이를 위해, 본 논문 전체에서 다음 정의를 사용한다.

Definition 1[10]: 특정 도메인 이름(domain name)의 웹사이트 D에 포함된 웹페이지의 집합을 $V=\{v_i|i\in\mathbb{N}\}$라고 하고, 웹페이지 간의 연결정보를 나타내는 하이퍼링크(hyperlink)의 집합을 $E=\{<v_i,v_j>|i,j\in\mathbb{N}\}$라고 하자. 여기서 $\mathbb{N}$은 자연수 집합이다. 그러면 웹사이트 D의 구조는 방향성 그래프 $G_D(V,E)$로 표현된다. 즉, 웹페이지 $v_1\in V$에서 웹페이지 $v_2\in V$로의 하이퍼링크를 통한 연결정보는 $<v_1,v_2>\in E$ 로 표현한다.

2.1 웹페이지 분야 적합도

웹페이지의 분야 적합도를 구하기 위해 웹페이지에 포함된 용어를 다음과 같이 정의한다.

Definition 2: 웹페이지 $v\in V$에 포함된 중복을 제거한 용어 t의 집합을 $T_v=\{t_i|i\in[1,n],\ i,n\in\mathbb{N}\}$라 하고, t의 출현빈도를 f_t라 하자. 그리고 v에 포함된 t와 f_t의 쌍으로 구성된 집합을 $P_v=\{(t_i,f_{t_i})|i\in[1,n],\ i,n\in\mathbb{N}\}$로 정의한다.

Definition 2에서 n은 웹페이지에 포함된 중복을 제거한 용어의 개수를 의미한다. 다음은 Definition 2의 예를 보인다.

Example 1: 웹페이지 $v\in V$에 포함된 용어가 $\{t_1,t_2,t_3,t_2,t_3,t_3\}$라 하자. 여기서 중복을 제거한 용어 집합은 $T_v=\{t_1,t_2,t_3\}$이고, 원소의 개수는 $n=3$이다. 또한 v에

1) "웹페이지"는 인터넷에서 특정 URL을 통해 접근할 수 있는 텍트스 묶음으로 정의된다. 또한 "웹사이트"는 인터넷에서 IP 주소나 도메인 이름(domain name) 만으로 이루어진 URL을 통해 접근할 수 있는 웹페이지들의 집합으로 정의된다. 따라서 본 논문에서는 IP 주소나 도메인 이름으로 웹사이트에 접속 시 가장 먼저 접근하는 웹페이지("main webpage")를 "홈페이지"(homepage)라고 부르기로 한다.

서 각 용어의 출현 빈도가 각각 $f_{t_1}=1$, $f_{t_2}=2$, $f_{t_3}=3$ 이므로, $P_v=\{(t_1,1),(t_2,2),(t_3,3)\}$이다.

특정 분야의 정보가 글로 표현될 때에는 일반적으로 그 분야에 관련된 용어들이 자주 사용된다. 이러한 개념에 기반하여 본 연구에서는 문서의 분야 관련도를 판단하기 위해 문서에 출현하는 전체 용어들 중에서 분야 관련 용어의 상대적 비율을 고려한다. 따라서 특정 문서에 포함된 용어의 중요성을 측정하기 위한 척도로 사용되는 TF(Term Frequency)[14]의 개념을 확장하여 다음과 같이 "분야 용어의 출현 빈도에 기반한 웹페이지의 분야 적합도"를 측정한다.

Algorithm 1: T_v의 정의가 Definition 2와 같고, 분야 용어 집합을 $Dic=\{d_i|i\in\mathbb{N}\}$이라고 하자. 그리고 $a,b\in\mathbb{N}$, $a+b=n$ 일 때, 웹페이지 $v\in V$의 용어집합이 $T_v=T_v^t\cup T_v^d$, $T_v^t\cap T_v^d=\varnothing$, $T_v^t=\{t_i|i\in[1,a],i\in\mathbb{N}\}$, $T_v^d=\{d_i|i\in[1,b],i\in\mathbb{N}\}$이고, v에서의 용어와 용어 출현 빈도의 쌍 집합이 $P_v=\{(t_1,f_{t_1}),\cdots,(t_a,f_{t_a}),(d_1,f_{d_1}),\cdots,(d_b,f_{d_b})\}$라 하자. 그러면 분야 용어의 출현빈도에 기반한 웹페이지 v의 분야 적합도 $S_v^{tf}\in[0,1]$는 식 (1)과 같다.

$$S_v^{tf}=\frac{\sum_{i=1}^{b}f_{d_i}}{\left(\sum_{i=1}^{a}f_{t_i}+\sum_{i=1}^{b}f_{d_i}\right)} \tag{1}$$

다음은 Algorithm 1의 예를 보인다.

Example 2: 웹페이지 $v_1,v_2\in V$에서 $d_1,d_2\in Dic$ 일 때, $P_{v_1}=\{(t_1,1),(t_2,2),(t_3,3),(d_1,2),(d_2,2)\}$, $P_{v_2}=\{(t_1,1),(t_2,2),(t_3,2),(d_1,5)\}$ 라고 하자. 그러면 $S_{v_1}^{tf}$, $S_{v_2}^{tf}$는 다음과 같다.

$$S_{v_1}^{tf}=\frac{f_{d_1}+f_{d_2}}{f_{t_1}+f_{t_2}+f_{t_3}+f_{d_1}+f_{d_2}}=\frac{2+2}{1+2+3+2+2}=0.4$$

$$S_{v_2}^{tf}=\frac{f_{d_1}}{f_{t_1}+f_{t_2}+f_{t_3}+f_{d_1}}=\frac{5}{1+2+2+5}=0.5$$

즉, v_1,v_2에 출현하는 용어의 전체 개수는 10 개로 동일하다. 그러나 분야 용어의 출현 빈도는 v_2가 v_1보다 더 크므로, v_2의 분야 적합도가 v_1보다 더 크다.

Algorithm 1은 같은 분야 용어가 반복적으로 출현할 경우에도 문서의 분야 적합도가 높게 측정된다. 반면에 다양한 종류의 분야 용어를 포함하는 문서는 좀 더 폭넓은 분야 정보를 표현한다고 볼 수 있다. 따라서 본 연구에서는 문서에 출현하는 중복을 제외한 전체 용어 중에서 분야 관련 용어의 비율을 고려하여 "분야 용어의 다양성에 기반한 웹페이지의 분야 적합도"를 측정한다.

Algorithm 2: T_v의 정의가 Definition 2와 같고, T_v^t, T_v^d, Dic의 정의가 Algorithm 1과 같다고 하자. 그러면 분야 용어의 다양성에 기반한 웹페이지 $v\in V$의 분야 적합도 $S_v^{tv}\in[0,1]$는 식 (2)와 같다.

$$S_v^{tv}=\frac{b}{a+b} \tag{2}$$

다음은 Algorithm 2의 예를 보인다.

Example 3: 웹페이지 $v_1,v_2\in V$에서 $d_1,d_2\in Dic$ 일 때, P_{v_1}과 P_{v_2}가 Example 2와 같다고 하자. 그러면 $T_{v_1}=\{t_1,t_2,t_3,d_1,d_2\}$, $T_{v_1}^t=\{t_1,t_2,t_3\}$, $T_{v_1}^d=\{d_1,d_2\}$이고, $T_{v_2}=\{t_1,t_2,t_3,d_1\}$, $T_{v_2}^t=\{t_1,t_2,t_3\}$, $T_{v_2}^d=\{d_1\}$이다. 또한, $T_{v_1}^t,T_{v_1}^d,T_{v_2}^t,T_{v_2}^d$의 용어 개수를 각각 a_1,b_1,a_2,b_2라고 할 때, $S_{v_1}^{tv}$와 $S_{v_2}^{tv}$는 다음과 같다.

$$S_{v_1}^{tv}=\frac{b_1}{a_1+b_1}=\frac{2}{3+2}=0.4,\ S_{v_2}^{tv}=\frac{b_2}{a_2+b_2}=\frac{1}{3+1}=0.25$$

즉, v_1의 분야 용어의 출현 빈도는 v_2보다 낮으나 상대적으로 다양한 종류의 분야 용어를 포함하므로, v_1의 분야 적합도가 v_2보다 크다.

분야 용어의 출현 빈도와 다양성에 기반한 웹페이지 분야 적합도 측정 방법을 결합하여 다음과 같이 웹페이지의 최종 분야 적합도를 측정한다.

Algorithm 3: 웹페이지 $v\in V$에 대해 S_v^{tf}와 S_v^{tv}가 각각 식 (1), (2)와 같다고 하자. $\mathbb{R}$이 실수 집합이고, $\alpha,\beta\in\mathbb{R}$, $\alpha\in[0,1]$, $\beta\in[0,1]$, $\alpha+\beta=1$ 일 때, 웹페이지 v의 분야 적합도 $S_v\in[0,1]$는 식 (3)과 같다.

$$S_v=\alpha\times S_v^{tf}+\beta\times S_v^{tv} \tag{3}$$

Algorithm 3에서는 분야 용어의 출현 빈도 또는 다양성에 가중치를 부여하여 웹페이지의 최종 분야 적합도를 결정한다. 또한, $S_v^{tf}\in[0,1]$, $S_v^{tv}\in[0,1]$ 이고, 두 양의 가중치 α와 β가 $\alpha+\beta=1$의 관계이므로 $S_v\in[0,1]$이다. 즉, $\alpha=1$이면 S_v^{tf}만으로, $\beta=1$이면 S_v^{tv}만으로 웹페이지의 분야 적합도 S_v가 결정된다.

2.2 웹사이트 분야 적합도

사용자는 홈페이지로부터 하이퍼링크를 통한 단계적 접근 방법을 통해 웹사이트에 포함된 웹페이지에 접근할 수 있다. 또한, 일반 문서 분류와 달리 웹페이지간의 연결 정보를 나타내는 하이퍼링크는 웹페이지의 분류에 중요한 정보가 될 수 있다[1]. 따라서 본 연구에서는 "특정 분야에 관련된 웹사이트는 분야 정보를 많이 포함한 웹페이지들이 홈페이지에서 접근하기 쉽게 구성되어 있다"는 것을 전제로 하여 웹사이트의 구조정보를

웹사이트의 분류에 이용한다.

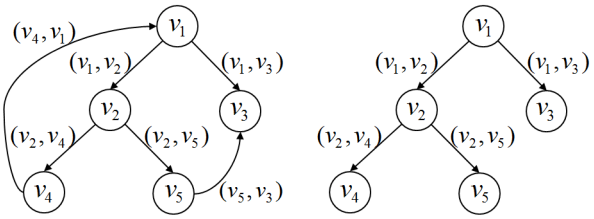
웹사이트는 Definition 1에서와 같이 방향성 그래프 구조로 표현되나 참고문헌 [12]에서와 같이 너비우선 신장트리(breath-first spanning tree) 형태로 변환함으로써 특정 웹페이지의 접근 용이성을 판단할 수 있다. 따라서 웹사이트의 분야 적합도 측정을 위해 다음과 같이 웹사이트의 구조를 변형한다.

Definition 3: 웹사이트 D의 구조가 Definition 1과같이 방향성 그래프 $G_D(V,E)$로 표현된다고 하자. 또한, D의 홈페이지 $v_h \in V$를 기준으로 너비우선 신장트리 구조(이하 "트리구조")로 표현한 방향성 그래프를 $TG_D(V,E_T)$라 정의한다. 여기서 $E_T \subseteq E$는 트리구조에서 웹페이지 간의 계층정보를 나타내는 하이퍼링크 집합이다.

Definition 4: 웹사이트 D의 트리구조 $TG_D(V,E_T)$가 Definition 3과 같다고 하자. D의 홈페이지 $v_h \in V$로부터 $v \in V$까지 접근하기 위해 거쳐야 하는 웹페이지 v_i, $v_j \in V$의 순서가 $v_h \rightarrow v_i \rightarrow \cdots \rightarrow v_j \rightarrow v$ 와 같을 때, 이들 웹페이지를 연결하는 하이퍼링크를 순서리스트로 나타내면 $L_v = (<v_h,v_i>, \cdots, <v_j,v>)$과 같다.

Definition 3에서 순서리스트 L의 크기는 트리구조에서 특정 노드의 레벨(level)로 볼 수 있다. 다음은 Definition 3과 4의 예를 보인다.

Example 4: 그림 1(a)와 같이 웹사이트 D가 $V=\{v_1, v_2, v_3, v_4, v_5\}$의 웹페이지들을 포함하고, 웹페이지들은 하이퍼링크 $E=\{<v_1,v_2>, <v_1,v_3>, <v_2,v_4>, <v_2,v_5>, <v_4,v_1>, <v_5,v_3>\}$로 연결되는 방향성 그래프 $G_D(V,E)$ 구조라 하자. v_1이 홈페이지일 때, $E_T=\{<v_1,v_2>, <v_1,v_3>, <v_2,v_4>, <v_2,v_5>\}$의 하이퍼링크를 이용하여 그림 1(b)와 같은 $TG_D(V,E_T)$를 생성한다. 또한, 홈페이지 v_1에서 웹페이지 v_4로의 접근을 위한 하이퍼링크의 순서리스트는 $L_{v_4} = (<v_1,v_2>, <v_2,v_4>)$이다.



(a)방향성 그래프 형태의 웹사이트 구조 (b)트리 형태의 웹사이트 구조

그림 1. 웹사이트 구조의 변형

Fig. 1. Transformation of an website structure

Definition 3과 4를 이용한 웹사이트의 분야 적합도 측정 방법은 다음과 같다.

Algorithm 4: 웹사이트 D의 트리구조 $TG_D(V,E_T)$가 Definition 3과 같고, 홈페이지 $v_h \in V$로부터 $v \in V$까지의 하이퍼링크 순서리스트 L_v이 Definition 4와 같다고 하자. 웹페이지 v의 분야 적합도 S_v가 Algorithm 3과 같을 때, 웹사이트 D의 분야 적합도 S_D는 식 (4)와 같다.

$$S_D = \frac{\sum_{v \in V} \left(\gamma^{l_v} \times S_v \right)}{m} \tag{4}$$

여기서 $\gamma \in \mathbb{R}$, $\gamma \in (0,1]$이고, l_v은 순서리스트 L_v에 포함된 하이퍼링크의 개수이며, m은 V에 포함된 웹페이지의 개수이다.

Example 5: 웹사이트 D_1과 D_2가 그림 1(b)와 같은 트리구조로써 상호 동일하고, 두 웹사이트에 포함된 웹페이지 집합 $V=\{v_1,v_2,v_3,v_4,v_5\}$에 대한 분야 적합도가 표 1과 같다고 하자. 즉, D_1에서 $S_{v_3}=0.24$, $S_{v_5}=0.515$이고 D_2에서 $S_{v_3}=0.515$, $S_{v_5}=0.24$이며, 나머지 웹페이지들의 분야 적합도는 서로 같다고 하자. 그러면 $\alpha=\beta=0.5$, $\gamma=0.9$일 때, S_{D_1}와 S_{D_2}는 다음과 같이 계산된다.

$$S_{D_1} = \frac{0.9^0 \times 0.325 + 0.9^1 \times 0.22 + 0.9^1 \times 0.24 + 0.9^2 \times 0.14 + 0.9^2 \times 0.515}{5} = 0.25391$$

$$S_{D_2} = \frac{0.9^0 \times 0.325 + 0.9^1 \times 0.22 + 0.9^1 \times 0.515 + 0.9^2 \times 0.14 + 0.9^2 \times 0.24}{5} = 0.25886$$

즉, D_1과 D_2에 포함된 웹페이지들의 평균 분야 적합도는 0.288로 동일하다. 그러나 D_2에서는 분야 적합도가 높은 웹페이지의 레벨이 D_1보다 상대적으로 낮으므로 D_1보다 D_2의 분야 적합도가 높게 계산된다. 즉, $\gamma<1$일 때, 분야 적합도가 큰 웹페이지의 레벨이 낮을수록 전체 웹사이트의 분야 적합도는 높아진다.

표 1. 웹페이지의 분야 적합도 예

Table 1. An example of the suitability of webpages

website	webpage	S_v^{tf}	S_v^{tv}	S_v	level
D_1	v_1	0.42	0.23	0.325	l_{v_1}=0
	v_2	0.26	0.18	0.220	l_{v_2}=1
	v_3	0.10	0.38	0.240	l_{v_3}=1
	v_4	0.17	0.11	0.140	l_{v_4}=2
	v_5	0.61	0.42	0.515	l_{v_5}=2
D_2	v_1	0.42	0.23	0.325	l_{v_1}=0
	v_2	0.26	0.18	0.220	l_{v_2}=1
	v_3	0.61	0.42	0.515	l_{v_3}=1
	v_4	0.17	0.11	0.140	l_{v_4}=2
	v_5	0.10	0.38	0.240	l_{v_5}=2

3. 웹사이트 분류기 구현

웹사이트에 포함된 웹페이지를 수집하고, 수집한 웹페이지로부터 웹사이트의 분야 적합도를 측정하기 위한 웹사이트 분류기의 작업 과정을 그림 2에서 보인다.

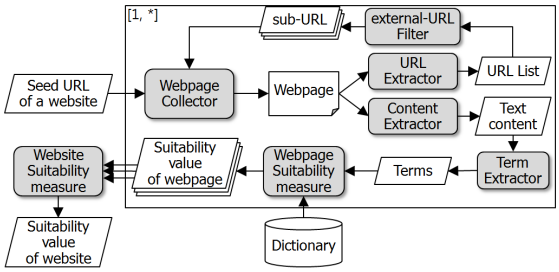


그림 2. 분야 웹사이트 분류를 위한 분야 적합도 평가 과정

Fig. 2. Process of the domain suitability evaluation for classification of medical field website

웹사이트에 포함된 웹페이지를 수집하기 위한 과정은 다음과 같다. 특정 웹사이트의 도메인 이름을 씨앗 URL(seed unified resource locator)로 하여 “Webpage Collector”를 통해 인터넷으로부터 웹페이지를 수집한다. 수집한 웹페이지로부터 “URL Extractor”를 이용하여 웹페이지에 포함된 하이퍼링크를 추출하고, “external-URL Filter”를 이용하여 추출한 하이퍼링크 중에서 외부 웹사이트의 웹페이지로 연결되는 하이퍼링크를 제거한다. 그리고 추출된 하이퍼링크를 이용하여 다른 웹페이지를 수집한다. 이러한 과정을 반복함으로써 웹사이트에 포함된 웹페이지들을 모두 수집한다.

웹사이트의 의료분야 적합도 측정은 다음의 과정을 따른다. “Content Extractor”를 이용하여 수집한 웹페이지에서 HTML과 script와 같은 태그를 제거함으로써 텍스트 형태의 데이터를 추출하고, 추출한 텍스트 데이터로부터 “Term Extractor”를 이용하여 용어를 추출한다. 또한 “Dictionary”에 포함된 분야 용어와 웹페이지에서 추출한 용어를 비교하여 웹페이지의 분야 적합도를 측정한다. 그리고 웹페이지에 대한 분야 적합도와 웹사이트의 구조 정보를 기반으로 웹사이트의 분야 적합도를 측정한다.

4. 실험 및 결과

제안한 웹사이트의 분야 적합도 측정 방법을 이용하여 의료분야의 웹사이트 분류 실험을 수행하였다. 실험을 위해 국내의 한 인터넷 포털사이트에서 제공하는 디렉토리 검색 서비스[2]의 의학관련 디렉토리[3]에 등록된 80 개의 웹사이트에서 4022 개의 웹페이지를 수집하였고, 다른 분야의 디렉토리에 등록된 127 개의 웹사이트에서 7176 개의 웹페이지를 수집하였다. 웹페이지 수집 과정에서 (1)외국어 사이트, (2)독립도메인(master domain name)의 웹사이트에 종속되어 운영되는 종속도메인(slave domain name)[4]의 웹사이트, 그리고 (3)수집한 웹페이지의 개수가 20 개 미만인 웹사이트는 실험 대상에서 제외하였다. 또한, 웹사이트에 포함된 전체 웹페이지의 양은 웹사이트마다 각기 다르며, 게시판을 포함할 경우에는 많은 하이퍼링크가 존재할 수 있다. 따라서 본 실험에서는 웹사이트마다 최대 100 개의 웹페이지를 수집하였다.

웹페이지의 텍스트 데이터로부터 한국어 명사를 추출하기 위해 한국어 형태소 분석기[15]를 이용하였고, 의학용어 표제어 사전[16]을 참고하여 의학용어 사전을 구축하였다. 의학용어 사전에 등록된 의학용어 중에는 “정보”(information), “이상”(abnormality, anomaly), “검색”(screening), “발생”(development), “생각”(thinking) 등과 같이 비 의료분야에서도 흔히 사용되는 범용어(common word)가 다수 포함되어 있다. 따라서 의학용어 표제어 사전에 등록된 4216 개의 의학용어 중 253 개의 범용어를 수작업으로 제거하고, 총 3963 개의 의학용어로 의학용어 사전을 구성하였다.

학습에 기반한 웹사이트 분류와 달리, 제안한 웹사이트의 분야 적합도는 이진 분류를 위한 기준이 없기 때문에 웹사이트의 이진분류(binary classification) 척도로 사용하기에는 적합하지 않다. 따라서 본 실험에서는 웹사이트의 분야 적합도를 기준으로 웹사이트들의 순위를 결정한 후, 상위 80 위에 포함되는 웹사이트 중에서 의학분야 웹사이트의 개수를 측정하여 분류 정확률을 계산하였다. 또한, 웹사이트의 분야 적합도는 α, β, γ 의 가중치에 따라 측정값이 달라질 수 있으므로, 가중치 값을 변화하면서 실험을 수행하였다.

그림 3은 가중치의 변화에 따른 웹사이트의 분류 정확률을 나타낸다. 가중치가 $\alpha=0.6$, $\beta=0.4$, $\gamma=0.9$ 또는 $\alpha\in[0.3,0.8]$, $\beta\in[0.2,0.7]$, $\gamma=0.6$ 일 때, 가장 높은 82.5 %의 웹사이트 분류 정확률을 보임을 알 수 있다. 그에 반해, $\alpha=1$, $\beta=0$ 이거나 $\alpha=0$, $\beta=1$ 와 같이, 용어 출현 빈도와 관련된 한 가지 인자를 제외하는 경우에는 웹사이트 분류 정확률이 상대적으로 낮아짐을 알 수 있다. 또한, $\gamma=1$ 일 때의 최대 웹사이트 분류 정확률은 81.3 %로 $\gamma<1$ 일 때보다 낮음을 알 수 있다. 따라서 웹사이트의 분야 적합도 측정을 위해 제안한 세 가지 인자(분야 용어의 출현 빈도, 분야 용어의 다양성, 웹사이트의 구조) 모두가 웹페이지 분류에 유용한 정보로 이용될 수 있음을 확인할 수 있었다.

그림 4는 실험에서 가장 큰 웹사이트 분류 정확률을 보인 가중치 $\alpha=0.6$, $\beta=0.4$, $\gamma=0.9$를 이용하여 웹사이트 분야 적합도를 계산하고, 이를 기준으로 웹사이트에 순위를 측정한 결과를 보인다. 그림에서 보듯이, 상위 80 위 까지는 총 66 개의 의학분야 웹사이트가 차지하고 있고, 그 중 58 위 까지는 의학분야 웹사이트가 차지하고 있음을 확인할 수 있었다. 분야 적합도가 낮은 의학분야 웹사이트를 분석한 결과, 성형외과나 치과

2) http://directory.search.daum.net/site_list.daum?dirseq=209767

3) [생활,건강>의료포털], [생활,건강>간호], [생활,건강>응급처치], [학문,사전>의학>의학잡지], [학문,사전>의학>단체], [학문,사전>의학>연구소,연구실]

4) 예를 들면, “www.daum.net” 이 독립도메인이라면, “blog.daum.net” 은 종속도메인이다.

와 같은 개인병원에서 운영하는 웹사이트나 의학분야 단체 관련 사이트가 대부분이었다. 이들 사이트에서는 상업적 목적을 위해 이미지와 플래시를 많이 사용하여 웹페이지가 텍스트 데이터를 많이 포함하지 않거나, 의학 정보 보다는 단체의 운영과 관련된 정보를 많이 포함하고 있었다.

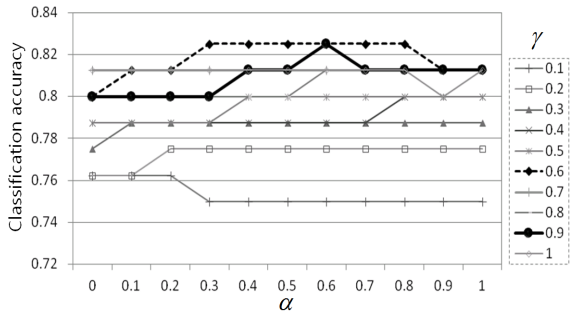


그림 3. 적합도 계산 가중치에 따른 웹사이트 분류 정확률

Fig. 3. Accuracy of website classification according to weights in the suitability measure

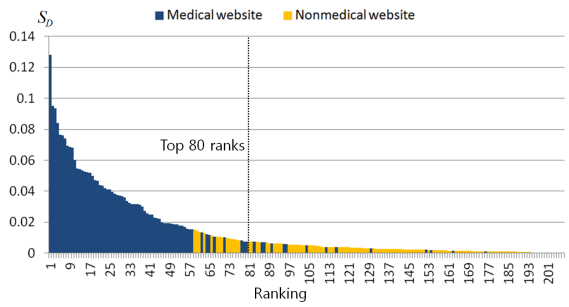


그림 4. 의료 분야 웹사이트 적합도 순위

Fig. 4. Ranking of websites' suitability on medical field

5. 결론

본 연구에서는 분야 용어의 출현 빈도와 웹사이트의 구조에 기반하여 웹사이트의 분야 적합도를 측정하는 방법을 제안하였고, 제안한 방법을 이용하여 의학분야의 한국어 웹사이트 분류 실험을 수행함으로써 제안한 방법의 효용성을 확인하였다. 기존의 학습 기반의 문서 분류 방법과 달리, 제안한 방법에서는 웹페이지에 포함된 의학용어의 출현 빈도와 웹페이지와 홈페이지 사이에 연결된 하이퍼링크의 개수만을 고려함으로써 웹사이트의 분류를 위한 웹사이트의 분야 적합도를 간단한 연산방법으로 측정하였다.

본 연구에서 수행한 실험에서는 HTML에서의 "<a>" 앨리먼트의 "href"의 속성 값으로 지정된 URL만을 수집하였고, script에서 사용된 하이퍼링크의 정보는 사용하지 않았다. 또한, 웹사이트에 포함된 웹페이지의 수를 제한하여 실험을 수행하였다. 이 문제를 개선함으로써 더 정확한 실험 결과를 얻을 수 있을 것으로 생각한다. 그리고 제안한 방법에서 측정한 분야 적합도는 하나의 웹사이트가 다수의 디렉토리에 분류될 수 있는 다중분류(multi classification)와 특정 디렉토리와의 관련성을 나타낼 수 있는 약분류(soft classification)에 이용될 수 있다. 따라서 다양한 분야의 세분화된 사전을 이용함으로써 다중 분류가 가능한 웹사이트 분류기를 설계할 수 있을 것으로 생각한다. 또한, 제안한 웹사이트 분류 방법에서는 규칙에 기반한 웹페이지의 분야 적합도를 계산하였으나, 기존의 학습에 기반한 문서분류 기법과 제안한 웹사이트 분류 기법을 융합하는 방법에 관한 연구를 수행함으로써 좀 더 정확한 웹사이트 분류 결과를 얻을 수 있을 것으로 판단한다. 향후에는 본 연구를 확장하여 한국어 웹사이트뿐만 아니라 외국어 웹사이트의 분류를 위한 연구가 수행될 필요가 있다.

References

[1] X. Qi and B.D. Davison, "Web Page Classification: Features and algorithms," *ACM Computing Surveys*, vol. 41, pp. 1-31, 2009.

[2] S. Chakrabarti, B. van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," *In Proceeding of the 8th International Converence on World Wide Web*, pp. 1623-1640, 1999.

[3] D. Mladenic, "Turning Yahoo into an automatic Web-page classifier," *In Proceedings of the European Conference on Artificial Intelligence*, pp. 473-474, 1998.

[4] S.S. Lee, "Korean Document Classification Using Extended Vector Space Mode," *KIPS Transactions: PartB*, vol. 18-B, no. 2, pp. 93-108, 2011.

[5] C. Li, D.R. Byun, and S.C. Park, "BPNN Algorithm with SVD Technique for Korean Document categorization," *Journal of the Korea Industrial Information System Society*, vol. 15, no. 2, pp. 49-57, 2010.

[6] W.H. Lee, S.J. Chung, and D.U. An, "Harmful Document Classification Using the Harmful Word Filtering and SVM," *KIPS Transactions: PartB*, vol. 16-B, no. 2, pp. 85-92, 2009.

[7] D.-H. Park, W.-S. Choi, H.-J. Kim, and S.-L. Lee, "Web Document Classification Based on Hangeul Morpheme and Keyword Analyses," *KIPS Transactions: PartD*, vol. 19-D, no. 4, pp. 263-270, 2012.

[8] N. Kim and J. Park, "Personal Information Detection by Using Naïve Bayes Methodology," *Journal of Intelligence and Information Systems*, vol. 18, no. 1, pp. 91-107, 2012.

[9] K.S. Ko, M.G. Hwang, P.K. Kim, and C.H. Lee, "Semantic Topic Selection Method of Document for Classification," *The Journal of the Korean*

Institute of Information and Communication Engineering, vol. 11, no. 1, pp. 163-172, 2007.
[10] M. Ester, H.-P. Kriegel, and M. Schubert, "Web Site Mining: A new way to spot Competitors, Customers and Suppliers in the World Wide Web," *In Proceedings of the 8th ACM SIGKDD*, pp. 249-258, 2002.
[11] Y.H. Tian, T.J. Huang, W. Gao, J. Cheng, and P.B. Kang, "Two-Phase Web Site Classification Based on Hidden Markov Tree Models," *In Proceedings of the IEEE/WIC International Converence on Web Intelligence*, 2003.
[12] O.-W. Kwon and J.-H. Lee, "Text Categorization based on k-nearest Neighbor Approach for Web site Classification," Information Processing and Management, vol. 39, pp. 25-44, 2003.
[13] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer, "The Connectivity Sonar: Detecting Site Functionality by Structural Patterns," In Proceedings of the 14th ACM Conference on Hypertext and Hypermedia, pp. 38-47, 2003.
[14] G. Salton, E.A. Fox, and H. Wu, "Extended Boolean Information Retrieval," Communications of the ACM, vol. 26, no. 12, pp. 1022-1036, 1983.
[15] "Espresso POS-K Tagger", Available: http://air.changwon.ac.kr/blog/2012/01/04/esspreso-pos-tagger-for-korean, [Accessed: July 26, 2012]
[16] 지제근, *알기쉬운의학용어 풀이집*, 고려의학, 2004.

저 자 소 개

이인근(In Keun Lee)
2001년 : 영남대학교 재료금속공학(학사)
2004년 : 영남대학교 대학원 전기공학(석사)
2010년 : 영남대학교 대학원 전기공학(박사)
2010년～현재 : 경북대학교 의료정보원천기술연구소 수석연구원

관심분야 : 의료정보표준, 개인건강기록, 온톨로지, 지능시스템
E-mail : inkeunlee@gmail.com

김화선(Hwa Sun Kim)
2003년 : 인제대학교 컴퓨터공학(석사)
2007년 : 경북대학교 의료정보학(박사)
2009년～2011년 : 경북대학교 의료정보학과 연구교수
2011년～현재 : 대구한의대학교 IT의료산업학과 교수

관심분야 : XML기반 병원정보시스템, 객체지향방법론 기반 CDA 및 RMI 개발, 임상표준용어코드
E-mail : pulala@paran.com

조 훈(Hune Cho)
1980년 : 서울대 수학과(학사)
2004년 : 미국 남캐롤라이나대학 전산학(석사)
2010년 : 미국 유타주립대학 의료정보학(박사)
1994년～1999년 : 아주대학교 의과대학 조교수
1999년～현재 : 경북대학교 의료정보학과 교수

관심분야 : 병원정보시스템, 온톨로지, 적정보상체계, HL7
Phone : +82-53-420-4899
E-mail : hunecho@knu.ac.kr