Project Name Focused Web Spider

Project Group Sercan DAĞDAŞ
Assist.Prof.Banu DİRİ (Advisor-2009)
Göksel BİRİCİK (Co-Advisor)

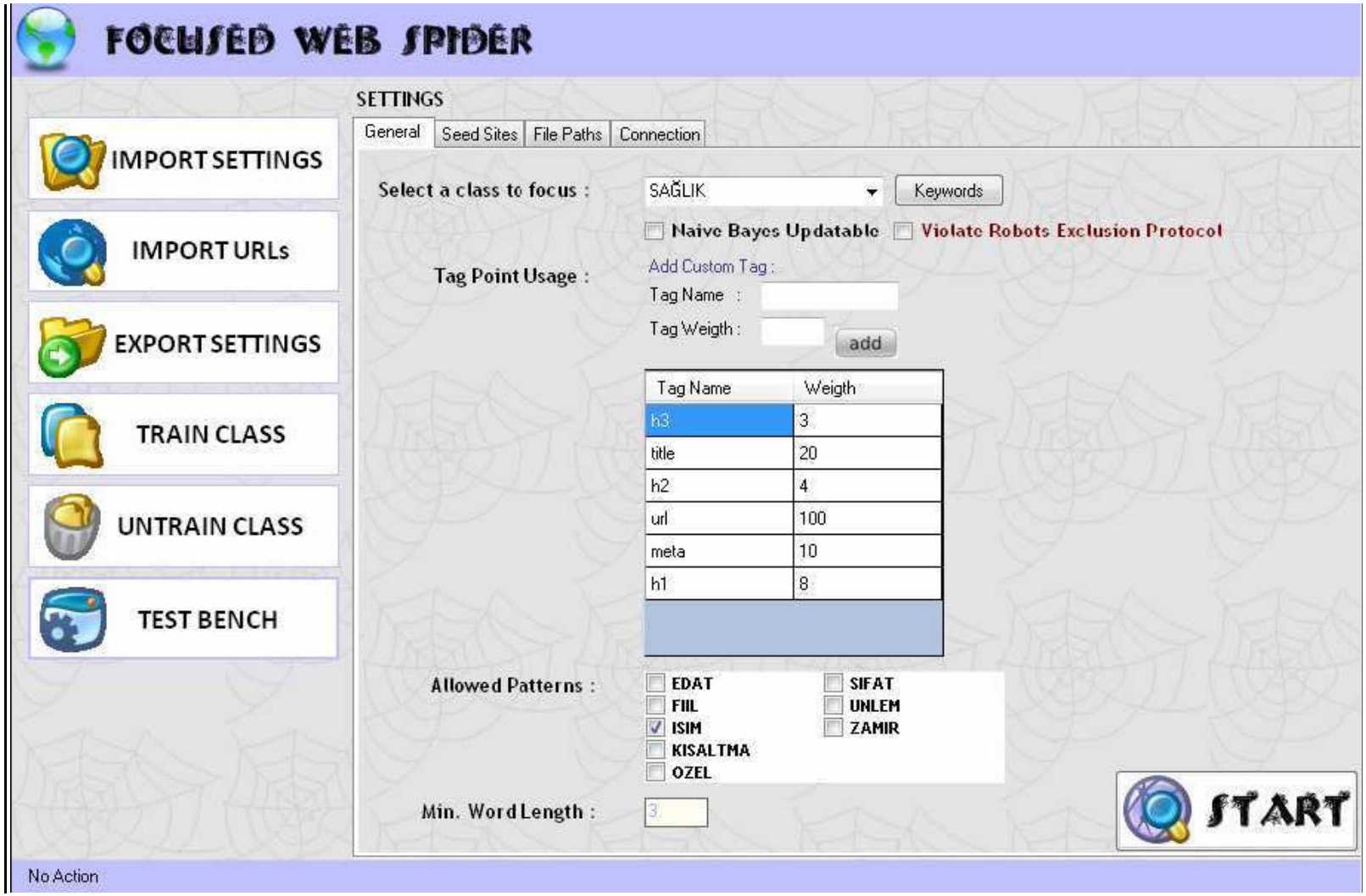
Abstract
The search engines that we use in our daily life often, searches the sites that are related with the searched words, with programs called crawlers, spiders or web robots. After that, these downloaded web pages are indexed in order to return results in very short time. “Web Spiders” are web page download automizing applications. They try to download entire site without obeying any rule. ”Focused Web Spider” download a content that is only related to a category or a subject. In this way, search engines do not browse unuseful pages. So, the results return not only fast but also more accurately. Moreover, browsing Internet in this way, enables optimum usage of network and disc resources.
First, we give a set of starting urls to “Focused Web Spiders” as we do in standard “Web Spiders”. But, it does not download the entire site. If the web page content is related with the required category this page is added to queue. After that, the same procedure applies to web pages in queue. Various algorithms (e.g. Naive Bayes Text Classification) are being used in order to classify a given web page. Moreover, some fields (such as html tags) may be more important than the others. And this is another aspect we should consider.

Özet (In Turkish)
Günlük hayatımızda sıkça kullandığımız arama motorları, aranılan kelime ile ilgili siteleri “crawler” veya diğer adlarıyla “spider”, “web robot” adı verilen uygulamalar ile tararlar. Daha sonra bu sayfalar indekslenerek sonuçların kısa sürede dönmesi sağlanır. “Web Spider”lar sitelerin indirilmesini otomatik olarak yapan uygulamalardır. Hiç bir şart olmadan tüm siteyi indirmeye çalışır. “Focused Web Spider”lar ise belli bir kategoriye, konuya odaklanarak bu konuyla ilgili içerikleri indirirler. Böylece arama motorları bu sayfaları kullanarak gereksiz sayfaları taramazlar. Hem hızlı, hem daha uygun sonuçlar döndürülmüş olur. İnternet’i bu sekilde taramak ayrıca ağ ve disk kaynaklarının çok verimli kullanılmasını da sağlar. “Focused Web Spider”lara taraması için standard “Web Spider” larda olduğu gibi, ilk önce bir başlangıç url seti verilir. Fakat onlar gibi tüm site indirilmez. Eğer içerik, ilgilenilen kategoriye ait bir sayfa ise bu bir kuyruk dizisine eklenir. Daha sonra aynı işlem bu kuyruk dizisi üzerindeki adresler için yapılır.

Verilen bir sayfayı ilgili kategori bazında sınıflandırabilmek için çeşitli algoritmalar (Ör: Naive Bayes Text Classification) kullanılır. Ayrıca bir web sayfasında bazı alanlar (html tag’leri gibi) diğer alanlardan daha önemli bir konuma sahip olabilir. Buna bağlı olarak da sayfanın ilgili kategoride yer alması için bir kriter daha belirlenmiş olur.

Requirements
Application will be written with C#.NET language, so application will run only Windows-based Operating systems. This project will be developed on Windows XP Operating System. This program will be getting following inputs from user:

- A set of seed urls
- Keyword set related topic
- Number of sites will be downloading
- Speed rate of download
- Depth of Web Graph


FOCUSED WEB SPIDER
IMPORT SETTINGS
IMPORT URLs
EXPORT SETTINGS
TRAIN CLASS
UNTRAIN CLASS
TEST BENCH
SETTINGS
General
Seed Sites
File Paths
Connection
Select a class to focus :
SAĞLIK
Keywords
Naive Bayes Updatable
Violate Robots Exclusion Protocol
Tag Point Usage :
Add Custom Tag :
Tag Name :
Tag Weigth :
add
Tag Name
Weigth
h3
3
title
20
h2
4
url
100
meta
10
h1
8
Allowed Patterns :
EDAT
FIIL
ISIM
KISALTMA
OZEL
SIFAT
UNLEM
ZAMIR
Min. Word Length :
3
START
No Action

Train your own class or improve an existing one !

Class Name :

Path to Documents :

Min. Word Length :

Allowed Patterns :

- [ ] EDAT
- [ ] FIIL
- [x] ISIM
- [ ] KISALTMA
- [ ] OZEL
- [ ] SIFAT
- [ ] UNLEM
- [ ] ZAMIR

Add Custom Tag :

Tag Point Usage :

Tag Name :

Tag Weigth :

add

Tag Name	Weigth
h3	3
title	20
h2	4
h1	8
meta	10

TRAIN CLASS

First Thread :

http://www.microsoft.com/turkiye/kampanya/outlook07.mspx saved on 15:05:11:937
http://www.microsoft.com/tr/tr/default.aspx saved on 15:05:05:371

Second Thread :

http://www.tebkobikulup.com/firma.aspx?fid=71 saved on 15:05:12:480

Id	Status	Web Page URL	ALIŞVERİŞ	BASIN VE YAYIN	BİLGİSAYAR	BİLİM	EĞLENCE VE YAŞAM	EKONOMİ VE İŞ DÜNYASI	KAYNAKLAR	KÜLTÜR VE SANAT	OYUN	SAĞLIK
7		http://www.microsoft.com/turkiye/haberler	0	100	59,4	56,9	39,5	28,1	56,8	38,3	61,2	41,5
6		http://entertainment.tr.msn.com/sifre.aspx	44,1	100	79,3	0	46,8	62,6	22,4	72	45	48,5
5		http://www.tebkobikulup.com/firma.aspx?fid=71 (NOT TRAINED)	84,3	88,9	100	79	85,3	97,2	75,3	85,2	0	77,7
4		http://www.microsoft.com/turkiye/kampanya/outlook07.mspx (NOT TRAINED)	88,4	84,7	100	78,3	85,2	90,3	84,2	72,5	0	82,7
3		http://www.microsoft.com/worldwide	0	0	88,3	33,1	26,2	40,3	96,7	38,8	32,7	30,7
2		http://www.microsoft.com/library/toolbar/3.0/sitemap/tr-tr.mspx ROBOTS.TXT : Page not downloaded	Robots.txt Protocol									
1		http://www.microsoft.com/tr/tr/default.aspx (NOT TRAINED)	0	29,4	100	2	36,5	44,2	27,2	45,5	10,8	23,6