Retrieving similar websites and web pages

BA Thesis (*Afstudeerscriptie*)

written by

Jevon van Dijk
(born December 13th, 1985 in Heemstede, Netherlands)

under the supervision of **Victor de Boer** and **Maarten van Someren**, submitted in partial fulfillment of the requirements for the degree of

BA Kunstmatige Intelligentie

at the *Universiteit van Amsterdam*.

Abstract

Similar web pages are pages that are about the same topic and of the same type. Sites about soccer clubs are related with all the soccer websites, but only similar with other soccer clubs. The goal of this research is to find an approach which, based on the textual content, can find similar pages given a page. The method used for this approach is a twofold method. The first task is trying to find similar websites, the second task is to find similar pages on those similar websites. Similarity is based on the textual content and use keyword extraction to identify the main topics. The tf*idf measure is used to identify the best keywords. For calculating the similarity between two pages the Cosine Similarity measure is used. The conceived approach gives some satisfying results for finding related websites and similar pages, finding similar websites is a difficult task. The conclusion of this research is that it is hard to find similar websites only based on the textual content. Finally, there are some suggestions given for further improvements.

Contents

1 Introduction

An important step in the design phase of a new website is to look at similar web sites. One of the reasons for the webmaster, the person who fills the website with content, is that he gets an overview from what topics can be found on that kind of websites and pages. Searching for similar websites can be time consuming. Therefore it would be useful to have a system to discover similar web sites automatically. In this thesis we discuss an implementation for such a program. We investigate several features from websites that can be used to compare different websites and web pages. This tool can also be used by the webmaster to compare changes on websites over some time. This useful because the web always changes, a page from two years old could be out of date. Using this tool makes it possible for the webmaster to compare his pages with other pages with the change to see new features for his pages.

This research topic is part of the Information Retrieval (IR) research area. More specifically, our research is an example of the of the text retrieval subtask of IR. The idea is to extract the relevant information from the given website(s) and calculate the similarity based on the textual content. In this research we use the Google engine for doing IR. The results, getting from the search engine, are processed and be re-ranked to determine which websites are more similar.

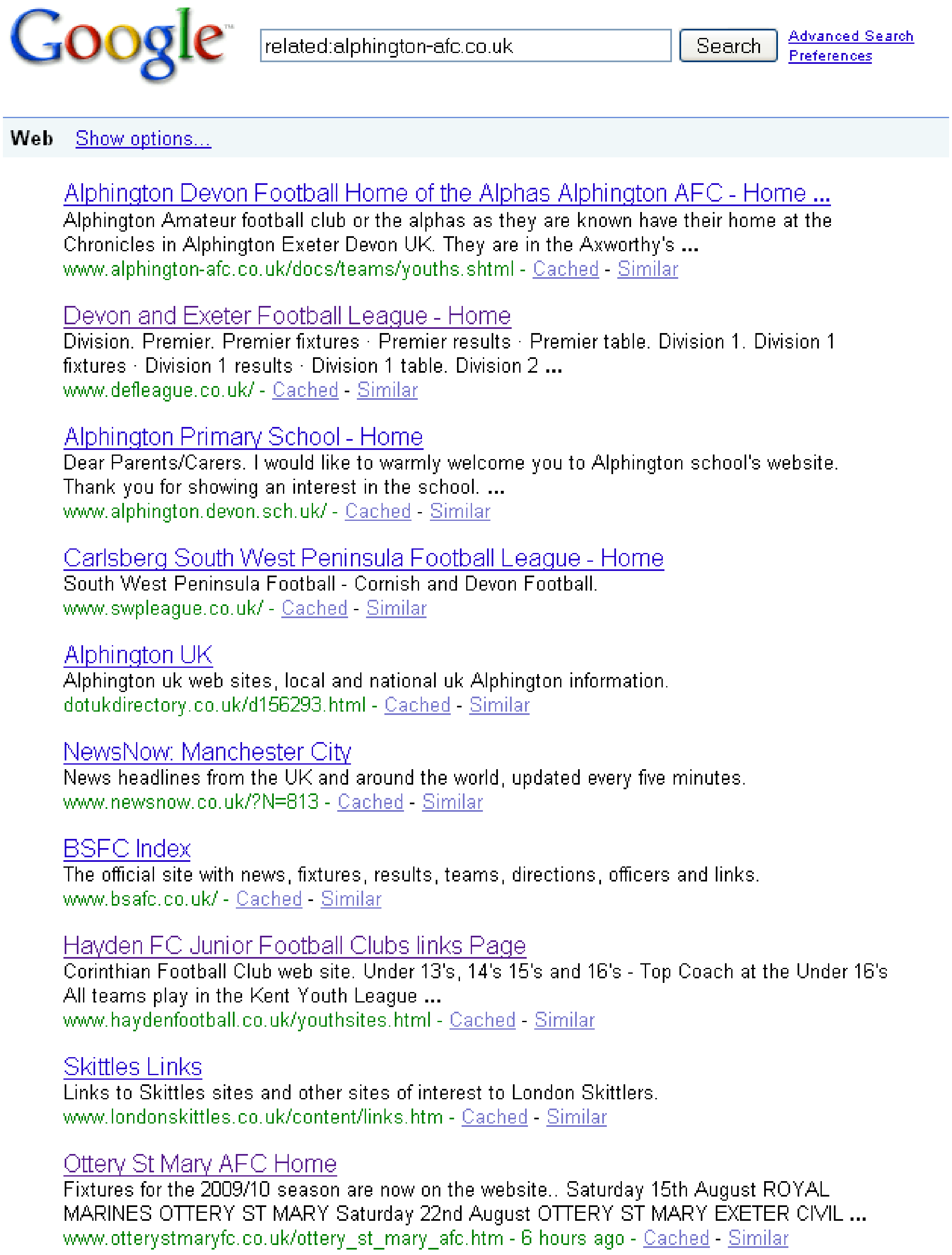
1.1 Similarity

This research have some interesting points, first of all Google [1] tries to do the same thing with there 'Similar pages' operation (related: search), but the results are not satisfying enough the way we define similarity. The definition of similarity in this case is that the websites are about the same topic and same type of website (e.g. all soccer club websites are similar to each other. Other soccer websites are not similar to soccer clubs, only related). So, a website is similar with another website when it contains the same type of information and also acts like the same type. The soccer clubs is there for the members of the club, to inform them about the club and teams. The reasons behind the websites are different. Google Related returns in the most cases only related websites and not related pages. If you search related websites for your local soccer club, you will find a lot of different results about soccer in general and just a few about soccer clubs. In figure 1 the top 10 results for the Alphington AFC (soccer club) are shown. When looking at the results, most of them are related based on the topic soccer (results 2, 4 and 6) or related with the city Alphington (results 3 and 5). The results the user want, are only those results which refers to soccer clubs, like results 8 and 10 from the results. So, this research is about retrieving similar websites and similar pages based on textual content. Similar means, website/pages of the same type with the same kind of information. The type of a website is important to say if a page is similar.

1.2 SiteGuide

The idea for this research is originally founded during the SiteGuide [2] project. In that project an application was developed that returns for a list of related websites returning a list of topics that be found on these websites. One disadvantage of the SiteGuide project is that those related websites need to be added manually. The approach in this research tries to deal with this problem because this research searches for those websites. The output in this research, a list of similar websites, could be used as input for the SiteGuide project. SiteGuide can make a model from this website. This also the reason why this research is so relevant with the SiteGuide project.

In the SiteGuide project [2] it is possible, given a number of manually added URL's of websites, to create a model for a new website structure. The results in that project were promising and gave opportunities for more advanced applications. By looking at some different criteria on the given pages the SiteGuide system decides if a page is a member of a certain cluster [3]. Each cluster contains pages (or part of pages or combination of multiple pages) with the same topics. It is important to notice that in the SiteGuide project



Alphington Devon Football Home of the Alphas Alphington AFC - Home ...
Alphington Amateur football club or the alphas as they are known have their home at the Chronicles in Alphington Exeter Devon UK. They are in the Axworthy's ...
www.alphington-afc.co.uk/docs/teams/youths.shtml - Cached - Similar

Devon and Exeter Football League - Home
Division. Premier. Premier fixtures · Premier results · Premier table. Division 1. Division 1 fixtures · Division 1 results · Division 1 table. Division 2 ...
www.defleague.co.uk/ - Cached - Similar

Alphington Primary School - Home
Dear Parents/Carers. I would like to warmly welcome you to Alphington school's website. Thank you for showing an interest in the school. ...
www.alphington.devon.sch.uk/ - Cached - Similar

Carlsberg South West Peninsula Football League - Home
South West Peninsula Football - Cornish and Devon Football.
www.swpleague.co.uk/ - Cached - Similar

Alphington UK
Alphington uk web sites, local and national uk Alphington information.
dotukdirectory.co.uk/d156293.html - Cached - Similar

NewsNow: Manchester City
News headlines from the UK and around the world, updated every five minutes.
www.newsnow.co.uk/?N=813 - Cached - Similar

BSFC Index
The official site with news, fixtures, results, teams, directions, officers and links.
www.bsafc.co.uk/ - Cached - Similar

Hayden FC Junior Football Clubs links Page
Corinthian Football Club web site. Under 13's, 14's 15's and 16's - Top Coach at the Under 16's All teams play in the Kent Youth League ...
www.haydenfootball.co.uk/youthsites.html - Cached - Similar

Skittles Links
Links to Skittles sites and other sites of interest to London Skittlers.
www.londonskittles.co.uk/content/links.htm - Cached - Similar

Ottery St Mary AFC Home
Fixtures for the 2009/10 season are now on the website.. Saturday 15th August ROYAL MARINES OTTERY ST MARY Saturday 22nd August OTTERY ST MARY EXETER CIVIL ...
www.otterystmaryfc.co.uk/ottery_st_mary_afc.htm - 6 hours ago - Cached - Similar

Figure 1: Google related results for Alphington AFC.

there is a difference between topics and pages. A page can contain several topics but also a topic can be divided over several pages.

These are the different criteria used in the SiteGuide project:

- **Textual content**, very useful to extract the main topic of a page;
- **Anchor text**, the text of a hyperlink can describe possibly the content of the page where it is linked to;
- **Pagetitle**, describes shortly the content of the pages (or website);
- **URL**, the url contains the filename. Often the filename contains the main topic of the page content, e.g. contact.html contains most likely the contact information;
- **Link structure**, distribution of internal and external links to pages in the cluster;

The content of pages contains everything the pages are about and we can assume that it must be possible, based on the textual content, to determine similarity. Using the *cosine distance* measurement and using the *tf*idf* measure for defining keywords can give us information about the similarity between two pages. The research basically focuses on similarity based on the textual content, but also considers why similarity based only on textual content is not strong enough to distinguish type of websites and pages.

The taken approach is a twofold approach, the first task is to find similar websites. The second task is to find similar pages inside those websites. Most important of the two tasks (when looking to the research aspect) is the first task, this is because we already know that the second task is possible due the SiteGuide project.

2 The problem

The investigation is a twofold approach. The input for this approach is a website and some pages (from that website). The output is some similar websites and pages, similar to the input website and pages. The first task is to find similar websites. The second task is to find inside those similar websites the similar pages. The main reason for doing it in two steps is that it more likely to get better results, because we want similar pages for the same kind of websites. If you search for similar pages for your restaurant's contact page, you will probably compare that page with other contact pages from other restaurants. In some cases there are differences for certain pages on very different websites. E.g. the menu page of a luxury restaurant contains some different information that the menu of the local chips and fries stand.

An important note is to distinguish between the terms *related* and *similar*. Google uses both terms for the same operation, but actually the operation finds only related websites following our definitions, which are websites that are about the same topic and of the same type. For example they are all about soccer, the results contains websites about soccer clubs, sites with soccer news, sites with only results and rankings, but they all have the same relation because the main topic is soccer. Similar sites are the same type of websites, for example the input is a website of a soccer club, a list of similar pages contains only links to other soccer clubs. This difference is also the difficult part to determine if a website is similar or only related. Related is easier to determine.

The research question for the first task is ”How can related websites be found based on the content of a given website?” The research is focused on efficient methods that can return a ranked list of related websites given a website based on the textual content. For the second task, the research question is ”How can we find related pages inside a given website based on the content of a given page?” Both tasks will consider a websites/web page as input for finding similar results.

3 Method

The approach is particular based on the textual content of the website and web pages. Which mean similarity is only based on the appearances of the words on the websites and pages. In this section we discuss how we collect candidate websites and how to determine if a websites or pages is similar. In Figure 2 the algorithm is systematically shown.

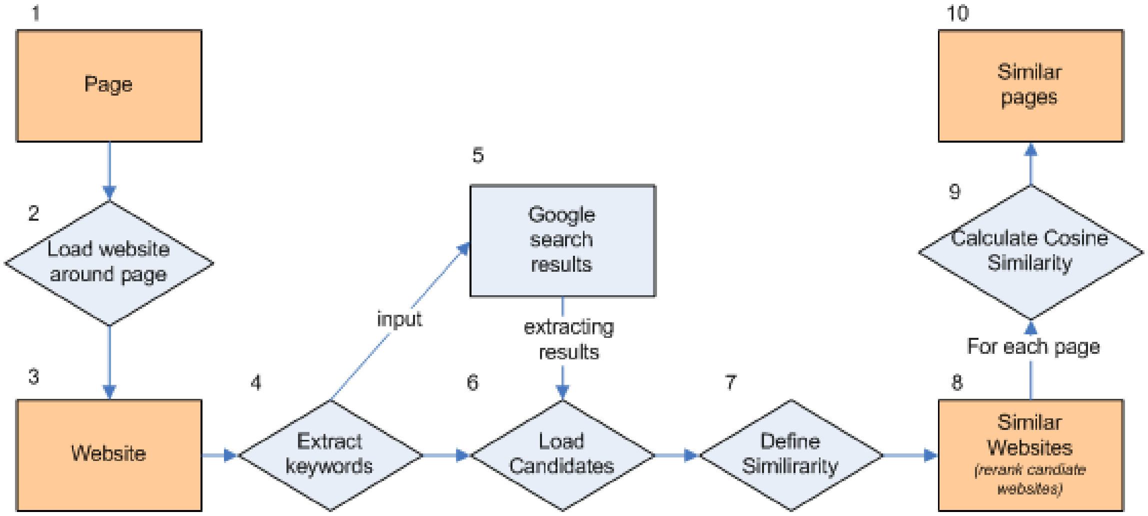


Figure 2: Overview of the algorithm for finding similar pages

The diagram in Figure 2 can be split up in four parts (orange rectangle boxes). The two parts (1 and 3) at the left are about the input website and the right parts (7 and 10) are the outcome of similar websites and pages. We can also split the parts horizontally, then the parts (1 and 10) at the top are about web pages and the parts at the bottom (3 and 7) are about websites. The grey line shows the relation of the four parts, which means we go from a page on a certain website, to similar websites and finally to similar web pages. In the following sections all the parts of the algorithm are addressed. Also the relations (or steps between the parts) are discussed.

The following points describe each step from the diagram from figure 2:

1. Input: URL of a page, for this page the system try to find similar pages.
2. Loading all pages from the website where the page of step 1 belongs to.
3. Create a website object, this object contains all information of the websites. E.g. list of words, links, pages, meta-tags, etc..
4. Using the words from the website for extracting keywords. Using the tf and idf measures.
5. Query keywords on the Google search engine, keywords based on step 4.
6. Parse the results from the Google search engine, store result information: Title, position, search terms, link. Load for every result the first page, the words of that page be stored. Words are needed to identify keywords for that page. Keywords are needed for calculating similarity with input page.
7. For all the candidate websites calculate the similarity with input website from step 3.
8. Rerank all websites based on their similarity value.

9. For all the pages of the related websites calculate the similarity with the input page (from 1).
10. Finally, rank all similar pages and return the best results.

3.1 Searching for similar websites

The system needs a collection of websites to find similar websites. Because we do not possess such collections, we use a search engine to extract a collection of websites (Step 5). TO get such a collection, using a search engine for finding websites is the easiest way. At the time of this thesis, Google is the world-leading search engine. It contains some interesting features to find possible similar websites. For example the search operations *link:* and *related:*. The websites we crawl we will call *candidate* websites. There are some useful operations to collect a set of candidate websites. To get useful results we use some different search operations to get a useful collection of candidate websites.

Google keyword search The given page contains a lot of text, from this text we extract the best keywords. Some similarity measures be used, like *term frequency* and *inverse document frequency*. Another option is to use the meta-tag keywords from the homepage of the website. From both keyword lists we remove the specific words (like the name of the restaurant) and use the best *n* words to search with Google. In section 3.1.1 more about extracting keywords. With the list of keywords we search for candidate websites.

Google related: search The *related:* search operation of Google is exactly what this research is about, only Google's similar pages can be improved. When you search for a local restaurant for related pages, you will find maybe some restaurants but also a lot of information about restaurants in general and information about things in the neighbourhood of that restaurant. The result we want is only those websites about the same type of local restaurants. Because there are some promising results it is useful to use these results and remove the not non-similar websites when handling these results. So, after the list of candidate website is analyzed the websites are ranked on similarity. Analyzing a website means post processing operations (e.g. keyword extraction) and calculating similarity with input website. The websites with the highest similarity value should be the most similar website.

Google link: search The *link:* search operation has the same advantages and disadvantages as the related operation. This operation returns a list of websites that has a link to the given website. The link tells already that there is a relation between both sites. The system must decide if that is the correct relation.

One point of discussion is the way we use search terms. Normally someone types the keywords in the search field in this way ”*bed breakfast hotel*.” How Google interprets this is not totally clear, possible it is a combination of *AND* and *OR* searching and it combines multiple words to one term. Possible it works better to say that we want an *AND* search operation: ”*bed AND breakfast AND hotel*.” But Google knows that the terms *bed and breakfast* can be seen as one term. In table 1 the differences are shown by using different search forms for the words *bed*, *breakfast* and *hotel*. The precision is compared with the results from query 1. Precision is defined as:

$$Precision = \frac{number\ of\ similar\ results}{total\ number\ of\ results} \tag{1}$$

Google takes ’bed AND breakfast’ as one term instead of the conjunction of two terms, this happens only in Query 3. The results show also a great difference in the total number of results. In this case there will always be millions of records and because we only parse a very small subset, and Google allows only the first 1000 results, it gives no problem. The 6th query, ”bed breakfast hotel,” executes the words as one term and the number of results (and precision) are much lower. This can be explained because the (exact) sequence of words appears little. Query 3 and 4 got the same results.

Query	Search terms	Number of results	Precision top 10
1.	bed breakfast hotel	24.500.000	1.0
2.	bed hotel breakfast	52.300.000	0.6
3.	bed AND breakfast AND hotel	34.000.000	0.5
4.	”bed” AND ”breakfast” AND ”hotel”	35.100.000	0.5
5.	bed AND hotel AND breakfast	50.300.000	0.6
6.	breakfast AND bed AND hotel	52.900.000	0.4
7.	”bed breakfast hotel”	183.000	0.2

Table 1: Using search terms affect the results

In this research there will only be focused on the normal search way (query 1) and *AND*-search (query 3), but it is important to know that the search results will be different for different keyword orders or by using Google’s binary search operators. In the following section (3.1.1) we discuss the way we extract keywords from websites and pages.

The corpus Most of our similarity measures are based on textual content, as explained in the Section 3. A corpus is used to define uniqueness of words, which is needed to define good keywords. This corpus contains for all words the term frequencies. The corpus used in the implementation of or system contains the 7728 most frequent English words. The used corpus is one found on the internet [1]. For an example of the corpus structure see table 2.

Word	Frequency
the	61847
of	29391
and	26817
a	21626
in	18214

Table 2: The corpus contains for 7728 words the word frequency

All the frequency are per million words, which means for example that the word *the* has a frequency of $\frac{61847}{1000000} = 0,061847$. These frequencies be used to define the keywords for a website and also to define if a word is too specific. The name of a restaurant can be too specific and when we search on the best keywords you do not want the restaurant’s name in the search query. So if the corpus contains a word, hopefully we can assume that that word is not specific enough; otherwise we delete that word from our query.

Beside the corpus, also a list of stopwords is used. These words are useful to delete uninformative words we want to leave out of our query, like *the*, *a(n)*, *are* and *be*. Totally we use 319 English words in our evaluation. In the following paragraph the effect on the results are pointed out and how dependent this approach is for language.

3.1.1 Extracting keywords

For a website or a page it is useful to define a set of the (best) keywords (Step 4). These keywords be used to search candidate sites in Google and to compare pages based on their content. There are several ways to get the keywords and the first one is the given set of keywords of the webmaster, these keywords can be found in the meta tags of the document and easily extracted. One disadvantage is that these keywords also contain

[1] Site from corpus: http://ucrel.lancs.ac.uk/bncfreq/flists.html

very specific words. For example take the homepage of The Guildford hotel [2]. The meta-keywords are:

guildford, hotel, hotels, surrey, b b, accommodation, guilford, guildfor, budget, cheap

Words like 'Guildford' are too specific. To remove those words we use a corpus (see also the evaluation section 4). If words do not exist in the corpus they will not be used. The corpus contains the most common words, so in this case Guildford is not one of them, so this word is removed from the set. For The Guildford hotel this results in the following set of keywords:

hotel, hotels, accommodation, budget, cheap

Another way is to define the keywords based on the frequency of words on the page or website. For each word we define the word frequency (equation 2). The $c(w, d)$ means how many times a word occurs in a document d, which can be a page or a total website. This number must be divided by the total number of occurrences of all documents. For a page this is the total number of pages inside the website. For a website is it the frequency stored in the corpus. The higher the normalized term frequency, the more important the word is for that document. An improvement of the results is to add manually one (or more) similar website(s), instead of only using the input website. In this case for each page we define a large set of keywords, but we take only the keywords that appear on both sites. This is a very efficient way to remove too specific keywords.

$$tf_w = \frac{c(w,d)}{\sum_i c(w,d_i)} \tag{2}$$

The *inverse document frequency* (equation 3) gives a value which indicates the general importance of a term and is an attempt measure to handle (near) stop words, which means that a word like "the" gets a low value because it is not very interesting. First we divide the total number of documents (for a page this is the number of all pages inside the website and for a website it is the word frequency from the corpus) by the number of documents where the word appears. From that number we take the logarithm to get the inverse document frequency.

$$idf_w = log\frac{|D|}{|d : w \in d|} \tag{3}$$

Finally, we multiply both the *term frequency* and the *inverse document frequency* (equation 4). The result is a measure which is used to get the best keywords.

$$tf \cdot idf_w = tf_w \times idf_w \tag{4}$$

An improvement to get better keywords is to add manually a similar page. It is then possible to define keywords that for both websites are relevant, too specific terms are filtered out.

3.1.2 Executing keyword queries

Now it is possible to define a set of keywords. With these keywords we can search to get candidate websites (Step 5). The following queries be executed to get the candidate websites:

- Best n meta-tag keywords

[2] The Guildford hotel: http://www.theguildford.co.uk/

- Extracted best n keywords from content
- Matching n keywords from input website and manually added similar website
- Google related search (related:*siteurl*)
- Google link search (link:*siteurl*)

For all these queries we parse the first n results. This results in a set of *number of queries* × *n* candidate sites. Due time reasons, mostly n is around the 10 results. The reasons that this operation is time expensive with large number of n, is because for every result it is needed to load the page behind that result. In this on-line approach for each similar search action we need to load around the 50 (when n is 10) pages. Loading a page is time expensive due slow web servers, large page or not available pages. Not available pages could happen when the web server is too busy. The 50 results, in the example above, can contain multiple results. This is because some query operations contain the same results. To avoid duplicate results it is needed to parse only unique results.

For each result from the queries, the link, title, text and position are stored. Also the page behind the link is loaded in the system. Due computational reasons we only take the first document and not the entire website. Taking the whole website will improve the results, because the first page may contain less or too specific information. Now each result be compared with the input website, the similarity value is set so it is possible to get a ranked list of candidate websites. Finally, after searching candidates and identify similar websites the result is a ranked list of similar pages based on their similarity value.

3.2 Searching for similar pages

To determine the similarity between a page and an input page from step 1, a number of statistical methods for ranking those pages are available. In this section we discuss the most interesting and useful methods to do this. For searching similar pages we use the retrieved similar websites from the previous section (3.1) or select manually a set of similar websites. From those websites we try to load all possible pages. In the page collection we try to find the most similar pages based on the textual content. The first step is to define the interesting keywords.

3.2.1 Similarity between pages

Similarity between two documents (Step 9) is calculated through the *cosine similarity* [4] measure that compares the occurrences of the words between the documents. The inputs for calculating the similarity are two pages: One the input page and the other a page from one of the similar websites. The equation 5 shows the formula to calculate the similarity between two documents, where Q is the query (in this case the words from a page) and this be compared with all candidate documents d_i. w_d is the *word frequency* for a word in a document d. The *word frequency* is already defined in the section *3.1.1 Defining keywords*. Above the division line the dot product is calculated and below the division line the Euclidian distance between the two word vectors. The result is the cosine distance which indicates how similar two documents are.

$$sim(q, d_i) = \frac{\sum_i w_q \cdot w_{d_i}}{\sqrt{\sum_i w_q^2}\sqrt{\sum_i w_{d_i}^2}} \tag{5}$$

For all the similar websites all pages are loaded (Step 2). This is done by looking at the first page (is the input url or the url from the Google results) of the websites and recursively following the internal links, until all pages will be found.

After parsing all the websites (defining keywords, etc.), the next step is to iterate over all pages (from the similar websites) and define the similarity based on the cosine distance between the input page (Step 1) and the candidate similar pages (Step 7). If all pages are parsed, the system can rank the results and give them back to the user (Step 10).

4 Evaluation

The evaluation is based on the precision of the results. For similar websites, only the top 10 of the results will be evaluated. Define manually how many websites are similar, this determines the precision. The number of 10 is chosen because we want not too many similar websites, so we return only the top 10. For similar pages the top 1 and top 4 are used for evaluation. Top 4 is chosen because it is only preferable when a similar pages is high ranked, so minimal in top 4 in the case the approach can not optimal determine the similar page.

Language dependency Using a corpus and stopword list could make our results language dependent. So, to evaluate the effect of these language models, we investigate the results of the system with a stopword list and without a stopword list. The tests are executed on five (1 input, 4 test) Dutch soccer sites. The input is a page which contains the information about the membership of the soccer club Alliance'22 [3]. The task is to find the same page on the other similar websites.

The test is to find a page which contains the information about the membership of that page. The results are interesting; one site gives us no page back because the page can not be parsed [4], on the other three websites without stopword list the similar pages are ranked lower then with the stopword list. See table 3 and Example Outputs 1 and 2 (Appendix A) for the results.

Website	With stop word list	Without stop word list
ado20.nl	2	3
dsov.nl	1	1
konhfc.nl	-	-
rch-voetbal.nl	1	2

Table 3: Ranking of a manually defined correct page

The stopword list improves the results on all three websites which contains similar pages. Here only the stopword list is tested, because we check here only on similarity between pages. In most languages stopword lists are available, so it is not a big issue to make this approach work in other languages.

Google related We compare our results with the results from Google Related. But first of all, the used approach differs in one important way from Google Related. Google Related only finds websites and no web pages, so on that point this approach is innovative and an improvement of Google Related. But it is more interesting to compare this on similar websites. An example to show that Google find only websites is to query a website *related:alphington-afc.co.uk* and a page *related:alphington-afc.co.uk/docs/teams/index.shtml*. Looking at the top 10 of the results, for both search operations the results are (exactly) the same.

[3] Input site: http://www.alliance22.nl/index.php?option=com_content&view=article&id=49&Itemid=18

[4] Reason is that the file is a PDF file. Because we only parse HTML-documents we can not define similarity

4.1 Test sets

To evaluate the approach there are two test set available on different domains. The first is a collection of hotels and the second test are amateur soccer clubs. These sites are used for input. So, for similar websites the links are used to find similar websites and for similar pages all the links are added manually to find inside these websites similar pages.

4.1.1 Test set: Hotels

This test set is based on five hotels in England. The sites are chosen manually and look relatively very similar. These pages also used in the SiteGuide project [2]. The task is to find similar websites given one of those websites and given one page on this website to find similar pages on the other websites. The five websites are:

1. The Guildford Hotel (www.theguildford.co.uk)
2. The Chandlery (www.thechandlerybandb.co.uk)
3. The Gatwick White House Hotel (www.gwhh.com)
4. Ye Old Talbot (www.yeoldetalbot.tablesir.com)
5. Masslink House (www.masslinkguesthouse.co.uk)

First we take a look at similar pages. In Example output 3 and 4 the results for this task shows us low precision scores. For Example 4 we manually added one website, which must lead to beter keywords because we use only keywords which are relevant for both websites. Example 3 got only one similar website (precision 0.1), adding a second website (Example 4) gives one extra homepage (precision 0.2). Comparing with Google related (precision 0.6) this is low. On both results, all pages are related to the keywords *hotel* (precision 1.0) and it is hard to distinguish between one website about a particular hotel and booking websites for hotels (which contains many hotels). One possible reason that Google got higher precision is the fact that we only parse the first page of a website. Parsing the whole website gives a better view of a site, which can lead to a better distinguish between hotels and other hotel related websites.

For the task of identifying a similar page, the results are more satisfying. For the Hotel example we search for the page where you can find the location of the hotel. This page contains mostly the address, route and map of the Hotel. For the four websites we found three times the similar pages on the first place, which means a precision of .75. If we look to the top 4 results, the precision is 1.0. The results can be found in the Appendix A, under Example output 5.

4.1.2 Test set: Amateur Soccer clubs

The second test contains ten amateur soccer clubs which plays in the English region Devon. In this test the same aspects will be tested like the Hotel set. The following ten amateur clubs websites are used for finding inside these websites a similar page and one website is used for finding similar websites.

1. Alphington AFC (alphington-afc.co.uk)
2. Buckland Athletic (www.bucklandathleticfc.co.uk)
3. Dartmouth United (www.dusc.ca)
4. Ivybridge Town FC (www.ivybridgefc.com)

5. Plymouth Parkway FC (www.plymouthparkway.com)
6. Plymstock United FC (www.plymstockutdfc.co.uk)
7. Stoke Gabriel AFC (www.stokegabrielafc.co.uk)
8. Teignmouth FC (www.teignmouthfc.co.uk)
9. Tiverton Town FC (www.tivertontownfc.com)
10. Totnes & Dartington SC FC (www.totnesdartington.com)

The first task is to find similar pages for the Alphington AFC soccer club. Where Google Related in the previous test, with Hotels, resulted in higher scores, the scores are now lower. In Example output 6, there are three similar websites (precision 0.3). Google Related found only two similar pages in the top 10 (precision 0.1).

Next, we evaluate the task of identifying similar pages inside the set of soccer websites. The goal is to try to find the history page for each club. The results can be found in Example output 7. The results are not as good as the hotel test set. Possibly the reason is that the location page is more often of the same type, the history pages could contain more different topics. Maybe some history pages are specific about some special events during the history of the club, which can make it difficult to define the exact topic. Only one site finds the similar page on position one. The top 4 results in a positive result in four websites, which means a precision of 0.4.

4.2 Overview results

Because the small test set and also the small number of test sets it is hard to prove which method is better, the approach discussed in this thesis or Google Related. The reason for the small test set is that it is difficult to find good test sets which contain everything what is needed for an good evaluation. A good test set requires a set of websites which can easily been parsed and analyze, but those test sets are hard to find. One advantages for this approach is that it can find similar pages and Google can not. With this approach it is also possible to use the results of Google Related to find similar pages inside those similar websites. The results difference a lot depending on the given input. Table 4 summarizes the results for similar websites and table 5 for similar pages.

Test set	Similar Websites	Google Related
Hotels	.1	.6
Hotels (manually added similar site)	.2	.6
Soccer clubs	.3	.1

Table 4: Precision for finding similar websites (top 10).

Test set	Top 1	Top 4
Hotels (Location page)	.75	1.0
Soccer clubs (History page)	.1	.4

Table 5: Precision for finding similar pages (top 1 and top 4).

5 Conclusion and discussion

The general conclusion is that it is possible to find, with positive results, related websites and similar pages. Based on the textual content it always finds related websites. Similar pages, depends on the quality of the input page. If that page contains enough information it is possible, with also positive results, to find similar pages. It is a lot harder to find similar websites based on textual content. The results show that keywords, based on textual content, are not strong enough for similarity for websites, but works fine for relatedness of websites and similarity for web pages. In this section the different good parts of this project are discussed.

Language models The use of language models, like a corpus and a stopword list is very useful to remove words with no extra meaning. So words like ”the”, ”a”, ”to”, ”be”, can easily be removed from the queries and keywords list. Also other specific words can be deleted by looking if they do not appear in the corpus, which can mean that the words are to specific, e.g. the name of the hotel or soccer club. A language model removes the noisy words from large bag of words. A disadvantage is that the corpus is not always available or large enough.

Topics The approach taken in this research is that every page is about one topic. Often this will work. But there are also websites that contain multiple topics on a page or topics divided over multiple pages. In the SiteGuide project [2] they use topics and they are able to classify better by using parts of pages or multiple pages together, probably the theory can be applied on this research as well to get a better idea what a page is about.

More data / Store data To get better results, it would be appreciating to use more data. At the moment, for each candidate website only the first page is loaded. This is because it is time expensive to load the whole website for each candidate. It is time expensive to load all the pages from a website, because the most websites contains a large quantity of pages. Another problem is to handle all different type of results, there is no uniform format for those websites, and so to handle all different type of pages is a difficult task. Another improvement is to store all information about websites and web pages locally, so there is no need to extract the keywords every time.

DOM-structure In this research, we used DOM-structure analysis on the web pages, which is for the most websites no problem. But there are also a lot of websites that can not be parsed, because those website are from another kind of website, which means is not a HTML-document. E.g. sites build up with Flash or JavaScript are almost impossible to parse. Some sites refer to PDF files, which also can not be parsed. Then there are also sites that use transformed addresses[5], which makes it hard to retrieve the site structure.

Website structure Why is it so hard to get only similar soccer and hotel websites and not related? The main problem is that based on textual content there is no difference, all the different types of sites contain mostly the same keywords. So, there must be other difference to distinguish those types. One of those things is to look to the structure of the website, often soccer clubs have the same page structures. Probably they contain a lot of team pages. These pages can not be found on websites about soccer news. Looking to those structures could be very useful, because it is a measure which can be used to differentiate different type of websites. This approach could be very useful in future work.

[5] This websites use for example the apache module *mod_rewrite*, which makes it possible to use '*fake*' links to access pages. E.g. site.com/index.php?page=home can be site.com/home. The web server converts site.com/home to the real filename.

Statistics about DOM Another useful point, is to look at statistical features about the DOM-structure of a HTML-page. Some type of websites contains a lot more images than other types, looking to the total number of images or the average number of images on a page could probably distinguish different types of websites. So, using these measures could improve the similarity between pages. Other measures could be the number of external links, the number of amount of money and more.

Finally It is important to understand that it is hard to meet high results, because the diversity of website types is so big. There are pages that could not be parsed because their website is not valid written or are not allowed to parse[6]. Because the standards of the web are not restricted and the diversity of options inside websites are enormous it is hard to find an approach that can handle all the parts of the web, there would always some things that can not be handled.

6 Acknowledgement

The first thanks go out to the supervisors of this research: Victor de Boer and Maarten van Someren. With the experience of both supervisors the level of this research is on a higher level. Also thanks to Bert Bredeweg and Wouter Beek for the general support and aid during the process of the Bachelor Thesis. The possibility to use the Google search engine makes the world much easier for this research. Finally, thanks to all the people from the UvA and outside the university for supporting me during the bachelor study and thesis.

7 Keywords

Cosine Similarity, DOM, Google, HTML, Information Retrieval, Inverse Document Frequency (idf), Relatedness, Semantic Web, Similarity, Term frequency (tf), tf*idf, Webcrawler, Webmaster tool, Webmaster

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine, in: Seventh international world-wide web conference (www 1998), brisbane, australia, 1998.

[2] V. de Boer, V. Hollink, and M.W. van Someren. Automatic web site authoring with siteguide, recent advances in intelligent information systems: Proceedings of the iis 2009 conference held in krakow, poland, 2009.

[3] V. Hollink, M.W. van Someren, and V. de Boer. Siteguide: An example-based approach to web site development assitance, in proceedings of the fifth international conference on web information systems and technologies (to appear), lisboa, portugal, 2009.

[4] Lillian Lee. Measures of distributional similarity. in proceedings of the 37th annual meeting of the association for computational linguistics on computational. *Linguistics*, pages 25–32, 1999.

[6] They could be excluded for web crawling with the use of a robots.txt file, which contains rules about web crawling.

Appendix A: Output results

In this appendix all the results are displayed which are discussed in this paper. In the text there are references to all this outputs. The results displayed with an asterisk before are the best results, in other words the results we want at the top of the list.

Example output 1 Trying to find pages about the "club membership," given one website and three similar websites.

```
Similar pages for
  http://www.alliance22.nl/index.php?option=com_content&view=article&id=49&Itemid=18
Similar website http://www.konhfc.nl/index.php loaded!
Similar website http://www.dsov.nl/index.php loaded!
Similar pages http://www.konhfc.nl/:
  1.332  http://www.konhfc.nl/jeugd_ouders_mededelingen.html
  1.182  http://www.konhfc.nl/sectie.php?sectie=Minis
  1.093  http://www.konhfc.nl/sectie.php?sectie=D-Sectie
  1.066  http://www.konhfc.nl/jeugd_arbiters_beleid.html
Similar pages http://www.rch-voetbal.nl/content/:
* 1.064  http://www.rch-voetbal.nl/content/index.php?
                      option=com_content&task=view&id=21&Itemid=63
  0.828  http://www.rch-voetbal.nl/content/index.php?
                      option=com_content&task=view&id=803&Itemid=2
  0.650  http://www.rch-voetbal.nl/content/index.php?language=en
  0.650  http://www.rch-voetbal.nl/content/index.php?id=2
Similar pages http://www.dsov.nl/:
* 1.461  http://www.dsov.nl//inlidmaatschap.htm
  1.208  http://www.dsov.nl/index.php?id=29
  1.208  http://www.dsov.nl//vbinformatie.htm
  1.153  http://www.dsov.nl//vbafgelastingen.htm
Similar pages http://www.ado20.nl/:
  1.554  http://www.ado20.nl/default.asp?page=cms&cms=1136&site=1
* 1.359  http://www.ado20.nl/default.asp?page=cms&cms=367&site=1
  1.315  http://www.ado20.nl/default.asp?page=cms&cms=532&site=1
  1.225  http://www.ado20.nl/default.asp?page=cms&cms=1713&site=1
```

There is no page found for www.konhfc.nl, the reason is that the page we are searching for on this website is not a HTML document. The file which contains the information about the membership is a PDF document, the system can not handle this type of files.

Example output 2 Trying to find pages about the "club membership," given one website and three similar websites, the same as Example output 1, but without a stopword list.

```
Similar pages for
  http://www.alliance22.nl/index.php?option=com_content&view=article&id=49&Itemid=18
Similar website http://www.rch-voetbal.nl/content/index.php loaded!
Similar website http://www.konhfc.nl/index.php loaded!
Similar website http://www.dsov.nl/index.php loaded!
Similar pages http://www.rch-voetbal.nl/content/:
  2.161  http://www.rch-voetbal.nl/content/index.php?option=com_content
         &task=view&id=803&Itemid=2
* 2.032  http://www.rch-voetbal.nl/content/index.php?option=com_content
         &task=view&id=21&Itemid=63
  1.915  http://www.rch-voetbal.nl/content/index.php?option=com_banners&
         task=click&bid=11
  1.827  http://www.rch-voetbal.nl/content/index.php?option=com_banners&
         task=click&bid=9
Similar pages http://www.konhfc.nl/:
  2.678  http://www.konhfc.nl/jeugd_ouders_mededelingen.html
  2.601  http://www.konhfc.nl/jeugd_arbiters_beleid.html
  2.393  http://www.konhfc.nl/nieuwspagina.php?&newsMsg=4817&newsCat=Nieuws
  2.324  http://www.konhfc.nl/sectie.php?sectie=Minis
Similar pages http://www.dsov.nl/:
* 2.673  http://www.dsov.nl//inlidmaatschap.htm
  2.670  http://www.dsov.nl//vbafgelastingen.htm
  2.670  http://www.dsov.nl/index.php?id=4
  2.528  http://www.dsov.nl/index.php?id=29
Similar website http://www.ado20.nl/default.asp?page=cms&cms=26&site=1 loaded!
Similar pages http://www.ado20.nl/:
2.765  http://www.ado20.nl/default.asp?page=cms&cms=532&site=1
2.724  http://www.ado20.nl/default.asp?page=cms&cms=1136&site=1
* 2.578  http://www.ado20.nl/default.asp?page=cms&cms=367&site=1
2.459  http://www.ado20.nl/default.asp?page=cms&cms=1713&site=1
```

Example output 3 Trying to find websites similar to the website of The Guildford Hotel.

```
Similar pages for http://www.theguildford.co.uk/Default.aspx
Keywords Meta-tags: hotel hotels accommodation budget cheap
Title keywords    : the hotel in hotel hotels in
Site keywords     : guildford hotel hotels room high bar
Matching keywords :
55 candidates handled, in 15.125 sec (avg: 0.275)
  0.  0.956  http://www.HotelSpecials.nl
  1.  0.865  http://www.theguildford.co.uk
  2.  0.774  http://www.hotel.com.au/Guildford-United-Kingdom/
  3.  0.768  http://www.world-stay.com/en/gb/surrey/guildford/
  4.  0.729  http://HotelsCombined.com/Guildford
  5.  0.585  http://Priceline.co.uk
  6.  0.575  http://realtravel.com/dh-17663-guildford_hotels
  7.  0.560  http://www.hotelscombined.com/hotel/the_guildford_hotel.htm
* 8.  0.550  http://www.prague-hotel-krystal.cz/
  9.  0.493  http://www.cheaperthanhotels.com.au/
```

Example output 4 Trying to find websites similar to the website of The Guildford Hotel, given manually one related website (Masslink House).

```
Similar pages for http://www.theguildford.co.uk/Default.aspx
Related website http://www.masslinkguesthouse.co.uk/default.aspx loaded!
Keywords Meta-tags: hotel hotels accommodation budget cheap
Title keywords    : the hotel in hotel hotels in
Matching keywords : hotel hotels room available rooms offer hire london stay comfortable
                    accommodation ensuite offers rates booking surrey gatwick +44 train
                    travelling guests walk cheap breakfast
92 candidates handled, in 49.0 sec (avg: 0.532608695652174)
  0.  0.881  http://www.dilos.com/hotel/4359
  1.  0.767  http://www.hotelclub.net/hotel.reservations/Guildford.htm
  2.  0.702  http://www.HotelCollect.com
  3.  0.650  http://www.europe-hotelrooms.com
  4.  0.644  http://www.tripadvisor.com/Hotels-g186390-Guildford_Surrey_England-Hotels.html
  5.  0.636  http://www.otshotels.com/
  6.  0.622  http://www.hotelsforever.com
* 7.  0.582  http://www.stokehouse.net/
  8.  0.572  http://www.priceline-europe.com
* 9.  0.518  http://www.ramadasurreyguildford.com/
```

Example output 5 Trying to find similar pages for The Guildford Hotel page which contains the location information.

```
http://www.theguildford.co.uk/guildford-hotel-directions.aspx (Location map restaurant)
```

```
=Start=
Similar website http://www.thechandlerybandb.co.uk/index.htm loaded!
Similar website http://www.gwhh.com/index.html loaded!
Similar website http://www.yeoldetalbot.tablesir.com/default.htm loaded!
Similar website http://www.masslinkguesthouse.co.uk/default.aspx loaded!

Similar pages http://www.thechandlerybandb.co.uk/:
* 0.480  http://www.thechandlerybandb.co.uk/mappage.html
  0.322  http://www.thechandlerybandb.co.uk/./index.htm#home
  0.322  http://www.thechandlerybandb.co.uk/index.htm#home
  0.322  http://www.thechandlerybandb.co.uk/index.htm
Similar pages http://www.gwhh.com/:
* 0.693  http://www.gwhh.com/location.html
  0.440  http://www.gwhh.com/parking-transfers.html
  0.402  http://www.gwhh.com/links.html
  0.358  http://www.gwhh.com/contact.html
Similar pages http://www.yeoldetalbot.tablesir.com/:
* 0.609  http://www.yeoldetalbot.tablesir.com/accommodation.htm
  0.294  http://www.yeoldetalbot.tablesir.com/becomeafriend.php
  0.293  http://www.yeoldetalbot.tablesir.com/default.htm
  0.288  http://www.yeoldetalbot.tablesir.com/conference.htm
Similar pages http://www.masslinkguesthouse.co.uk/:
  0.575  http://www.masslinkguesthouse.co.uk//default.aspx
  0.575  http://www.masslinkguesthouse.co.uk/default.aspx
  0.571  http://www.masslinkguesthouse.co.uk//pages/links.aspx
* 0.516  http://www.masslinkguesthouse.co.uk//pages/gatwick.aspx
```

Example output 6 Trying to find similar webste for Alphington AFC soccer club.

```
Search for related websites:
http://www.alphington-afc.co.uk/index.shtml

Similar website http://www.ivybridgefc.com/index.htm loaded!
Keywords Meta-tags:
Title keywords    : football home of the home page
Matching keywords : town youth league st division west club game united devon news team
                    player half goals ground
41 candidates handled, in 97.359 sec (avg: 2.374609756097561)
  0.  0.427  http://www.exeter.gov.uk/index.aspx?articleid=2313
  1.  0.356  http://dotukdirectory.co.uk/d156293.html
* 2.  0.273  http://www.stokegabriel.co.uk/
  3.  0.218  http://www.exeteryouthleague.co.uk/under12s-division1/
  4.  0.199  http://www.southern-football-league.co.uk/news/newsDec05.htm
* 5.  0.155  http://www.clubwebsite.co.uk/amesburytown/history.pl
* 6.  0.144  http://www.axminstertownfc.co.uk/news.asp?offset=40
  7.  0.132  http://genuki.cs.ncl.ac.uk/DEV/Alphington/
  8.  0.125  http://en.wikipedia.org/wiki/Alphington,_Devon
  9.  0.101  http://www.defleague.co.uk/Results/Senior_1
```

Example output 7 Trying to find similar history page, given the history page for the Alphington AFC soccer club.

```
Similar pages for http://www.alphington-afc.co.uk/docs/club/history.shtml
Similar website http://www.bucklandathleticfc.co.uk/index.html loaded!
Similar website http://www.dusc.ca/default.asp loaded!
Similar website http://www.ivybridgefc.com/index.htm loaded!
Similar website http://www.plymouthparkway.com/default.htm loaded!
Similar website http://www.plymstockutdfc.co.uk/index.html loaded!
Similar website http://www.teignmouthfc.co.uk/index.cgi loaded!
Similar website http://www.stokegabrielafc.co.uk/index.cgi loaded!
Similar website http://www.totnesdartington.com/index.php loaded!
Similar website http://www.tivertontownfc.com/Index.htm loaded!
Similar pages http://www.bucklandathleticfc.co.uk/:
  1.936  http://www.bucklandathleticfc.co.uk/swpl.htm
* 1.401  http://www.bucklandathleticfc.co.uk/club_history.htm
  1.306  http://www.bucklandathleticfc.co.uk/news.htm
  1.202  http://www.bucklandathleticfc.co.uk/t1reports.htm
Similar pages http://www.dusc.ca/:
  0.399  http://www.dusc.ca/default.asp?mn=1.2.27
  0.356  http://www.dusc.ca/default.asp?mn=1.23
  0.297  http://www.dusc.ca//default.asp?mn=1.2&y=2001&m=7
  0.297  http://www.dusc.ca//default.asp?mn=1.2&y=2001&m=5
Similar pages http://www.ivybridgefc.com/:
  1.010  http://www.ivybridgefc.com/archivenews.htm
  0.938  http://www.ivybridgefc.com/new/home.htm
  0.810  http://www.ivybridgefc.com/1stresults.htm
  0.625  http://www.ivybridgefc.com/1stfixtures.htm
Similar pages http://www.plymouthparkway.com/:
  1.154  http://www.plymouthparkway.com/swpl/fixRes.html
  1.086  http://www.plymouthparkway.com/default.htm
  1.068  http://www.plymouthparkway.com/o35/fixRes.html
  1.068  http://www.plymouthparkway.com/u9b/fixRes.html
Similar pages http://www.plymstockutdfc.co.uk/:
  0.839  http://www.plymstockutdfc.co.uk/5.html
  0.689  http://www.plymstockutdfc.co.uk/2.html
  0.313  http://www.plymstockutdfc.co.uk/4.html
  0.078  http://www.plymstockutdfc.co.uk/3.html
Similar pages http://www.teignmouthfc.co.uk/:
  0.652  http://www.teignmouthfc.co.uk/index.cgi
* 0.605  http://www.teignmouthfc.co.uk/history.pl
  0.435  http://www.teignmouthfc.co.uk/useful_links.pl
  0.430  http://www.teignmouthfc.co.uk/findus.pl
Similar pages http://www.stokegabrielafc.co.uk/:
* 0.869  http://www.stokegabrielafc.co.uk/history.pl
  0.825  http://www.stokegabrielafc.co.uk/index.cgi
  0.755  http://www.stokegabrielafc.co.uk/clubdocuments.pl
  0.715  http://www.stokegabrielafc.co.uk/club_news.pl
Similar pages http://www.totnesdartington.com/:
  1.067  http://www.totnesdartington.com/index.php?option=com_content&
         view=category&layout=blog&id=11&Itemid=54
* 1.067  http://www.totnesdartington.com//index.php?option=com_content&
         view=category&layout=blog&id=11&Itemid=54
  0.985  http://www.totnesdartington.com//index.php?option=com_content&
         view=section&layout=blog&id=3&Itemid=2
  0.985  http://www.totnesdartington.com/index.php?option=com_content&
         view=section&layout=blog&id=3&Itemid=2
Similar pages http://www.tivertontownfc.com/:
  1.796  http://www.tivertontownfc.com/clubhonourspags.htm
  1.257  http://www.tivertontownfc.com/league0001page.htm
  1.236  http://www.tivertontownfc.com/fixturespage.htm
  1.202  http://www.tivertontownfc.com/league0102page.htm
```