Study of Near Duplicate Content: Identification of Categories Generating Maximum Duplicate URL in Results

Kavita Garg
Research Scholar
School of Computer and Information Sciences
MVN University, Palwal

Jayshankar Prasad, PhD
H.O.D
School of Computer and Information Sciences
MVN University, Palwal
India

Saba Hilal, PhD

ABSTRACT
The study of identification of near duplicate content involves identifying search categories which generate same URL in a query result. These categories are needed to be identified so that results can be improved by removing duplicate URL. Generating same URL in results irritates the user and it also decreases priority of other URL. These URL displayed on second or third page which user do not bother to open. Near duplicate content sometimes hides better results from the user and make the search results ineffective. There are many algorithms and procedures or filters to reduce the duplicity. But to reduce duplicity there is need to identify that duplicates. Which categories generate most duplicate results, in what form redundancy exists, which search engine generates these duplicate results and so on. This paper shows efforts to identify categories with maximum duplicates in term of same URL.

General Terms
Duplicate Urls Identification

Keywords
Keywords are your own designated keywords which can be used for easy location of the manuscript using any search engines.

1. INTRODUCTION
To identify duplicate Urls, this paper considers six search engines which are widely used and sample search categories are selected...

1.1.Search Engines
To identify different search categories, it is requiring selecting different search engines which will generate results for those categories. To extract data from web, user uses Search engines. Search engines act as mediator between user and the

web. User submits the query and gets the results in response to the Query. There are hundreds of search engines available

that can process the Query**.** Some well-known search engines are **Google, Yahoo, Bing, Yandex, cuil, Ask, dog pile, AltaVista, AOL, Gig blast, Lycos** etc. Of the many available search engines, some of the search engines (Google, Yahoo, Bing, Ask, AOL, and Dog pile) that are most commonly used are selected.

1.2.Sample Search Categories
Today whole world is accessible through Web. Each and every type of information is available on web. This information comes under different Search category. Like, if someone is interested in knowing about Sachin Tendulkar, then it comes under the category of players. Search category is defined as the domain about which user want to gain Information or knowledge. There are different types of users who are interested in surfing different search topic. Also there are thousands of search categories about which user can retrieve information. In order to select the few search categories out of thousands, different users are selected who visits the internet randomly. Today every age (from teen age to old age), every class () of user uses the internet. When the users of different age group are asked for "**what they search over internet**", then different users shown different area of interest. On the basis of interest of use around 40-44 search categories are selected. Some of these categories are further subcategorize. List of search categories along with the user is summarized here:

Table 1.Different type of users and search categories

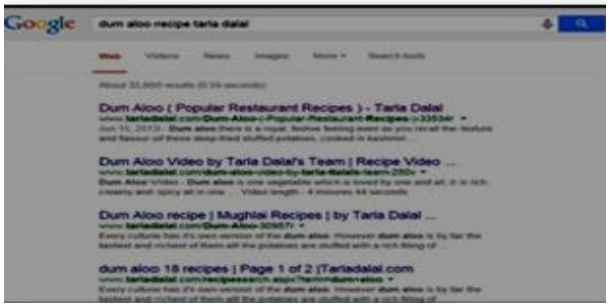
User	Search category
College students	Movies, games, perfumes, Mobile phones, Exam, Institutions.
Housewife	Song, cooking recipe, Diet plan, palmistry, Rhymes and poems, child psychology, home appliances, jewelry, acupressure.
Lecturer	Syllabus, Books, journals, codd rules
Working professionals	Hotels and resorts, Mobile banking, Car, Bikes, Bill payment, symptoms of diseases, passport services, schools, Trains route, Tourist place, shoes.
Women(age>50)	Eclipse , online religious TV serial , Medicine.
Retired person	Share market, News, banking, Powerful personalities.
Others	Polygraph test, Building and towers, ice cream, Awards and prizes, National games, freedom fighters, weather and climate.

The above table summarizes the categories used to find out duplicate content. Of the above mentioned categorize there are some categorize which are further subcategorize on the basis of the attribute of the category. For e.g. If someone is searching for a book then it may be possible that different user search it by different attribute. Someone search by book name, other search by publisher of book, someone can search by price of book.

1.3.Search Categories with Duplicate Results

In present time, 60% of the people uses internet in their daily life. Today, Web is a huge repository of information and it contains answers to every query. It is very easy to access or upload data over the web. Any person can easily upload the data over the web, because of which around 30% of the data is duplicate. Many algorithms have been developed to reduce this duplicity, so that storage space and serving cost can be saved but duplicity still exists.

Here is the sample result:



In the same way search was performed on 44 different categories and 6 different search engines (Google, yahoo, Bing, Ask, AOL, Dog pile) to find out categories which contains highly duplicated content. For each category there are five parameters, their values are obtained along with date and time. The five parameters are:

1) No. of link on page 1.
2) Similar link.
3) Dissimilar link.
4) Similar summary content
5) Search query string.

First parameter gives the value of how many of link on first page.

Similar link defines how many links are same or how many links directly or indirectly refers to the same page.

Dissimilar link gives the value of links which refer to different pages.

Similar summary content gives the brief idea about the link i.e. what content it may contain.

Search query string gives the idea that whether the link is relevant or not. If the query string is present in summary content then the link may be relevant but if it is not present in the summary content then the link is not relevant

Table2: Sample records of some categories on six different search engines.

Book of Computer Fundamental	Google	Yahoo	Bing	Ask	Aol	Dogpile
subcategory1(Author name) (P.K sinha)						
No. of link	10	10	10	10	10	10
Similar link	0	1	1	0	0	1
Dissimilar link	10	9	9	10	10	9
Similar summary content	9	6	4	3,2	6	7
Search query string	10	7	6	9	9	7
Date	27/4/15	27/4/15	27/4/15	27/4/15	27/4/15	27/4/15
Time	8:45PM	9:16PM	9:33PM	10:37PM	10:09PM	10:58 PM
subcategory2(Tutorials)						
No. of link	10	11	11	10	10	10
Similar link	4	2	2	2	2	2
Dissimilar link	6	9	9	8	8	8
Similar summary content	4	2		3	3	2,2
Search query string	8	5		7	7	3
Date	27/4/15	27/4/15	27/4/15	27/4/15	27/4/15	27/4/15
Time	8:58PM	9:25 PM	9:42PM	10:4pm	10:0pm	11:02pm
Subcategory3(price)						
No. of link	10	10	11	10	10	10
Similar link	1	2	4	3	0	1
Dissimilar link	9	8	7	7	10	9
Similar summary content	2	3	3	2	0	2

1.4.Identifying Categories which are Generating Duplicate URL

In this section, categories that mostly contain duplicate content are identified. There are seven parameters are linked with each category, each having its own significance and purpose. To identify duplicates in the category similar link parameter is used. This parameter gives the value of links that refer to the same URL or same page, directly or indirectly. When search is performed on 44 categories with 6 different search engines, the result shows that there are 8 such categories which contain most duplicated content. We are considering categories which are generating more than 5 similar links

Minimum value of similar link	**0**
Maximum value of similar link	**7**
Here threshold value is taken as	**>=5**

Table 3: Analysis of categories on the basis of similar link

Search categories	Maximum value of similar link
Shoes	
Subcategory1(shoes store) (woodland shoes store delhi)	5
Subcategory2(shoes online) (woodland mens formal shoes flipkart)	7
Cooking Recepie	
Subcategory1(Recpie by chef) (Dum aloo recpie by Tarla dalal)	7
Movies	
Subcategory2(Movies in PVR) (Movies in Srs cinema faridabad)	5
Weather and Climate	
Subcategory2(Weather forecast of day) (weather forecast of 1 May 2015 in delhi)	5
Rhymes and poems	
Subcategory(Rhyme name) (Two little dicky bird)	7
Banking services	
Interest rate of syndicate bank on fixed deposit	5
Ice cream	5
Amul ice cream variety	
Online Tv Serials	
Watch shri krishna online	6

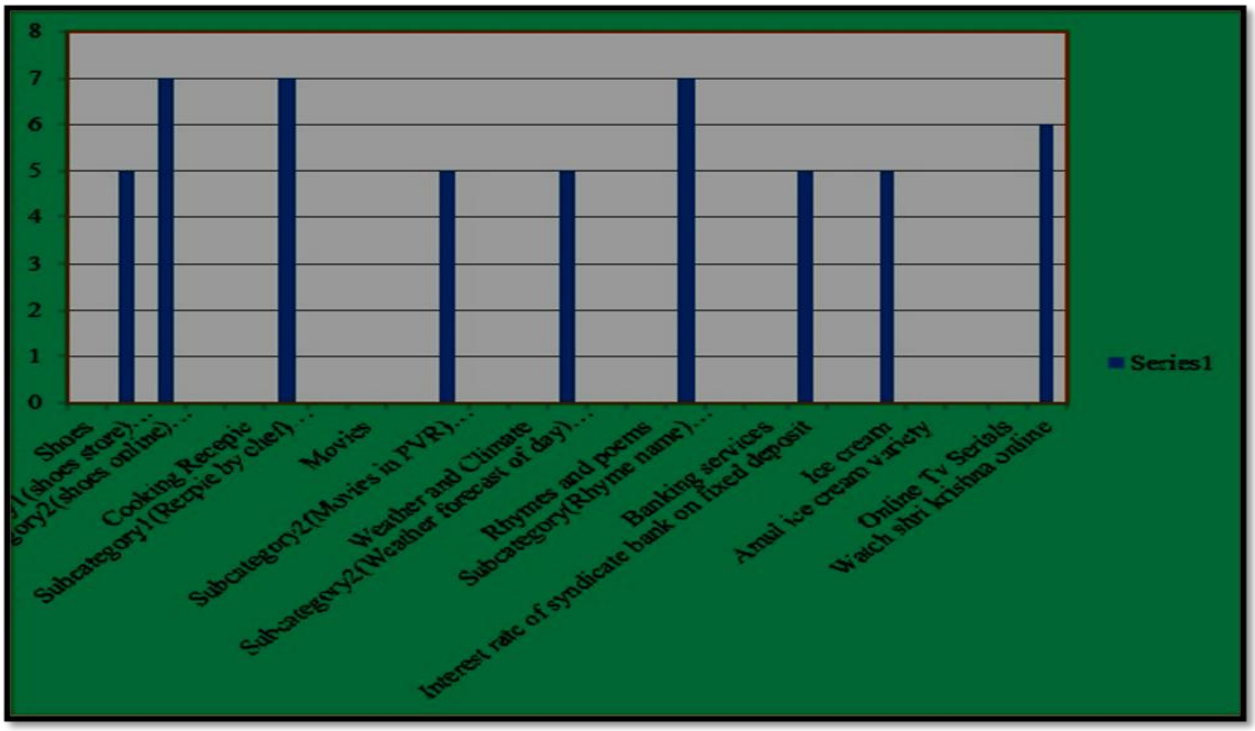


Fig1: Graphical Representation of categories generating maximum duplicate links.

2. RELATED SEARCHES

In [1], Hui Yang, Jamie Callan, Stuart Shulman give a new algorithm DURIAN(Duplicate Removal In lArge collection),to detect near duplicate .DURIAN uses a bag of words document representation, metadata and content structure of document to identify form letters and edited copies in public comment collection.

In [2], Sarah Weissman, Samet Ayhan, Joshua Bradley, and Jimmy Lin identify similar identical sentences in Wikipedia by implementing minhash to identify sentences with high Jaccard coefficients. Mainly focuses on identical sentence that determine of copy and paste and quality issues in Wikipedia.

In [3], Rekha V R, Resmy V R uses machine learning technique to identify the patterns of the URLs. URLs patterns are used to develop a framework for de-duplicating the web pages. The pair wise rules are generated from URL pairs and web crawlers apply these rules, to normalize the URLs. The normalized URLs are tokenized and pattern tree is constructed.

In [4], Manuel Egele, Christopher Kruegel, Engin Kirda developed a new approach that detects web spam pages in the result returned by search engines. First, it determines the importance of different page features to the ranking in search engine results. Based on this information, it develops a classification technique that uses important features to successfully distinguish spam sites from legitimate entries.

3. CONCLUSION

Today world is world of web. Every user searches query over the web according to their interest. There is no. of categories for which user search the query. This research paper selected 44 categories and studied those for finding duplicate URLs. Among 44 categories there are 8 categories which generate maximum duplicate URLs. The future work will concentrate on finding search engine which generate maximum duplicate links and will remove those duplicate links.

4. REFERENCES

[1]. H.Yang, J.Callan, S.Shulman (2006), “Next Steps in Near-Duplicate Detection for eRulemaking”, Proceedings of the international conference on Digital government research, pages 239-248.

[2]. S.Weissman, S.Ayhan, J.Bradley, and J.Lin (2015), “Identifying Duplicate and Contradictory Information in Wikipedia”, Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 57-60.

[3]. R.V R,et al,(2016) ,“Speeding up of Search Engine by Detection and Control of Duplicate Documents on the Web”, International Journal of Computer Science and Information Technologies,Vol.7 (2) , 637-642.

[4]. M.Egele, S.Barbara, E.Kirda(2011) ,“Removing Web Spam Links from Search Engine Results” Journal in Computer Virology, Vol.7(1), doi>10.1007/s11416-009-0132-6