ISSN:2229-6093

Tanvi Gupta et al,Int.J.Computer Techology & Applications,Vol 3 (1), 231-234

A NOVEL APPROACH TO DETECT THE NEAR-DUPLICATES BY REFINING PROVENANCE MATRIX

Tanvi Gupta

(Department of Computer Science)

Lingaya's University, Faridabad,India

Asst.Prof. Latha Banda

(Department of Computer Science)

Lingaya's University, Faridabad ,India

Abstract :- ***In this paper, the provenance matrix is refined to get more accuracy and efficiency in detecting near-duplicates by adding two more factors 'How' and 'Why' , as the performance of the web search depends on the search results having information without duplicates or redundancy . More redundancy leads to more time consume and more storage, that's why search engines try to avoid indexing of duplicates documents. Provenance model combines both the content-based and trust-based factors for classifying near-duplicates or original documents, as now a days, many of near-duplicates are from the distrusted websites .***

Keywords: near-duplicates, Provenance , distrusted , provenance matrix, trustworthiness

1. Introduction

Search Engines uses crawlers to gather information and stores it in database maintained at search engine side. For a given user's query the search engine searches in the local database and very quickly displays the results. But, In any web search environment there exist challenges when it comes to providing the user with most relevant, useful and trustworthy results, as mentioned below:

• The lack of semantics in web
• The enormous amount of near-duplicate documents
• The lack of emphasis on the trustworthiness aspect of documents

There are also many other factors that affect the performance of a web search.

Also, the new challenges are created by Web for information retrieval as the amount of information on the web and number of users using web growing rapidly.

A. Web Based Search Optimization using Provenance matrix:

One of the causes of increasing near-duplicates in web is that the ease with which one can access the data in web and the lack of semantics in near-duplicates detection techniques. It has also become extremely difficult to decide on the trustworthiness of such web documents when different versions/formats of the same content exist. Hence, the needs to bring in semantics say meaningful comparison in near-duplicates detection with the help of the 6W factors – Who *(has authored a document)*, What *(is the content of the document)*, When *(it has been made available)*, Where *(it is been available)*, Why *(the purpose of the document)*, How *(In what format it has been published/how it has been maintained)*[5]. This information can also be useful in calculating the trustworthiness of each document. A quantitative measure of how reliable that any arbitrary data is could be determined from the provenance information. This information can be useful in representative elimination during near-duplicate detection process and to calculate the trustworthiness of each document. The existing approaches of near-duplicates detection and elimination do not give much importance to the trustworthiness aspect of the content in documents retrieved through web search. Thus, Provenance based factors [1]. may be used for near-duplicates detection and elimination which provides the user with the most trustworthy results.

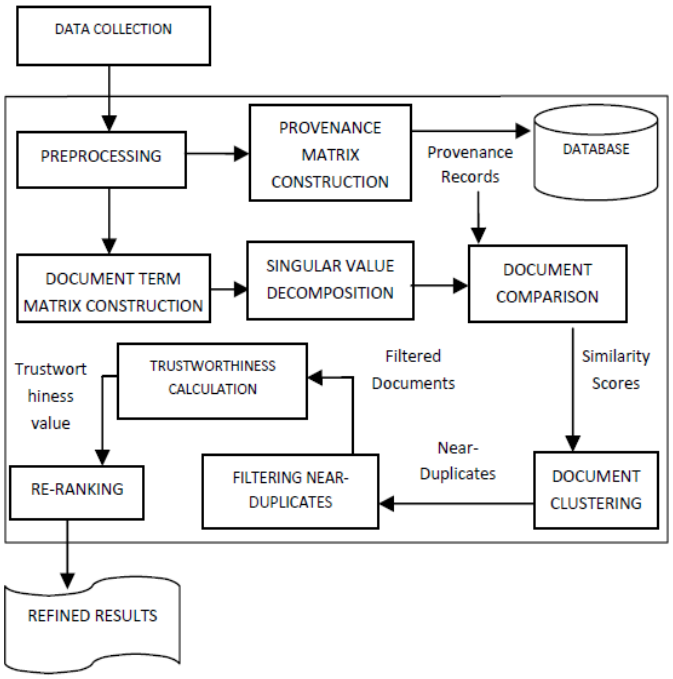

Fig.1 . Web Based Search Optimization using Provenance[1]

The architecture suggested by Y. Syed Mudhasir et.al[1] comprises of the following components: (i) Data collection (ii) Preprocessing (iii) Document Term Matrix(DTM) construction (iv) Provenance Matrix(PM) Construction (v) Database (vi) Singular Value Decomposition (vii) Document

Clustering based on similarity scores (viii) Filtering (ix) Re-ranking based on trustworthiness values.
Here, the provenance matrix is of the form:

factors/docs	doc 1	doc 2	doc3
Who	Andrew McCallum Kamal Nigam	Kamal Nigam Jing Luo	Lyle H. Ungar
Where	IEEE	ACM	IEEE
When	2000	2000	2004
What	Content of doc1	Content of doc2	Content of doc3

Fig.2 Provenance Matrix[1]

2. Related Work

A. Improving Relevance of Search Engine Results by Using Semantic Information from Wikipedia [6]: In order to meet the three types(transactional, informational and navigational) of searches, search engines basically use algorithmic analysis of the links between pages improved by a factor that depends on the number of occurrences of the keywords in the query and the order of these words on each web page returned as a result. For transactional and informational queries, the relevance of the results returned by the search engine may be improved by using semantic information about the query concepts when computing the order of the results presented to the user.

B. Semantic Approaches to detect near-duplicates: Salha Alzahrani et.al[2] gives a method on plagiarism detection using fuzzy semantic-based[2] string similarity approach. The algorithm was developed through four main stages. First is pre-processing which includes tokenization, stemming and stop words removing. Second is retrieving a list of candidate documents for each suspicious document using shingling and Jaccard coefficient. Suspicious documents are then compared sentence-wise with the associated candidate documents. This stage entails the computation of fuzzy degree of similarity that ranges between two edges: 0 for completely different sentences and 1 for exactly identical sentences. Two sentences are marked as similar (i.e. plagiarized) if they gain a fuzzy similarity score above a certain threshold. The last step is post-processing hereby consecutive sentences are joined to form single paragraphs/sections.
Krishnamurthy Koduvayur Viswanathan et.al[3] suggests a method to recognize that two Semantic Web documents[3] or graphs similar, and characterizing their differences is useful in many tasks, including retrieval, updating, version control and knowledge base editing. A number of text based similarity metrics are discussed as in that characterize the relation between Semantic Web graphs and evaluate metrics for three specific cases of similarity that have been identified: similarity in classes and properties used while differing only in literal content, difference only in base-URI, and versioning relationship.

3. Proposed Work

In Web Search Optimization based on Web Provenance matrix, the Data collected is in the form of html pages which is copyrighted by some company or a person, and also have information for server ie where it is store and also when this website was launched on internet. The Proposed work is the refined model of the provenance matrix.

Factors	Doc1	Doc2	Doc3
Who	Company or Person name of doc1 who has copyright of it.	Company or Person name of doc2 who has copyright of it.	Company or Person name of doc2 who has copyright of it.
When	Date or year of launch	Date or year of launch	Date or year of launch
Where	Server name	Server name	Server name
What	Content of doc1	Content of doc2	Content of doc3
Why	Title of the page or first heading in the body part of doc1	Title of the page or first heading in the body part of doc2	Title of the page or first heading in the body part of doc3
How	Format of doc1	Format of doc2	Format of doc3

Fig.3.Refined Provenance Matrix

This Refined Provenance Matrix includes:

(1) ho *(copyrighted by which company or person)*,
(2) hat *(is the content of the html page in body)*,
(3) When *(it has been made available(server name))*,
(4) Where *(it is been available)*,
(5) hy *(the purpose of the document)*,
(6) ow *(In what format it has been published/how it has been maintained)*

The focus is on two factors ie. 'Why' and 'How', 'Why' factor is define as the purpose of the html document which can be stated by either the title of the page or first heading of the body part and 'How' factor is define by the format of the html page . The html page having better format will be considered in web search optimization based on Provenance[1].

Better Format for the html pages is given by Web Authoring Standards[7] are as follows:

(1) Validation

All HTML documents shall be checked with an HTML validator.

(a) Documents should be validated at HTML level 2.0, 3.2 (Wilbur), or future specifications from W3C when available.

(b) Authors using a DTD other than the above shall include an appropriate DOCTYPE declaration. In the case of a DTD that is not standard or widely-known (eg those available from WebTechs validation service), the DTD itself shall be referenced in a comment within the document.

(c) Validation errors shall be noted by the author. Such notes may be delivered separately to the HTML documents (provided they are referenced by a comment in the source), and should describe the purpose of the invalid construct, together with its effect in several browsers including text-mode browsers (the comment "no effect" is acceptable). In the case of an invalid but established construct, a reference to an existing analysis is sufficient.

(2) HTML Headers

(a) All HTML documents shall include an appropriate TITLE .

(b) Documents may include other header elements, such as relational links, Stylesheets, Client-side scripts, and META elements.

(c) Documents which are a "front page" or other principal entry point for a system should include the following: KEYWORDS and DESCRIPTION meta elements for the benefit of Web indexers.

(3) Colours and Background Images

(a) Authors MAY use any legal markup to determine document colours, but should use RGB specifications to do so.

(b) Where colours are set by an author, they shall ensure a strong contrast between text and background. This implies light-on-dark or dark-on-light: colour contrasts are insufficient to cater for monochrome displays or colour-blind readers. Note that this implies that authors setting a text colour must also set the corresponding background, and vice versa.

(c) Background images (where used) should be small, and should be of a similar colour to the BGCOLOR specified.

(d) Conspicuous background images should be avoided in pages containing textual information.

(4) Images

(a) Images may be used to complement text, but should not be used to replace it. Examples of appropriate use are diagrams, graphs and geographic maps; inappropriate examples are passages of decorated text and imagemaps used to replace it.

(b) All images shall have ALT texts. Where appropriate, ALT=" " is acceptable. ALT texts for images which are also links shall be descriptive of the purpose of the link, and should be brief. "Home", "Next", "Previous", "Search" are examples of good ALT texts; "Click Here", "Home Icon", "Binocular Icon", "Back to XYZ Homepage" are irredeemably bad.

(c) ALT texts should not duplicate nearby document text.

(d) ALT texts for larger images (eg those above about 10Kb) SHOULD warn of their size - for example "Global Composite Image (29K)" (although it MAY be appropriate to omit this in cases where *any* non-blank ALT text would be obtrusive).

(e) ALT texts for imagemaps may direct readers to a separate text toolbar; otherwise they should be blank (ALT=" "). Where imagemaps are used, alternative means of navigation shall be made available to readers.

(f) Images should use height and width attributes, except as noted under Browser Compatibility below.

(5) Appropriate Use and Deprecated Tags

(a) <BLINK> shall not be used.

(b) <FONT> shall not be used in place of HTML headers <H1> - <H6>

(c) Emphasis tags such as <FONT>, <B> or <STRONG> SHOULD NOT be applied to extended passages. They are appropriate to words and phrases, and (exceptionally) as much as a complete paragraph of text.

(d) <H1>...</H1> should be used exactly once in an HTML page.

(6) Nonstandard/Proprietary Markup

(a) HTML pages MAY be "enhanced", they should not be dependent on proprietary markup. Specifically, all key functionality and information should be available to an HTML-compliant browser not supporting the "extension".

(7) Browser Compatibility

(a) HTML constructs which render a document difficult to read due to known defects in popular browsers should be avoided, regardless of the construct's validity in strict HTML.

(b) Use of < or > within a tag, in a construct such as <IMG SRC="forward.gif" ALT="=>"> risks breaking parsers and should be avoided.

(c) Comments should open with <!-- and close with -->. Use of "--" or ">" within these delimiters should be avoided.

(d) Height and Width attributes in images should be restricted to cases where neither the image itself nor the ALT texts are essential to the document. In particular, images which are navigation icons should not use height and width attributes, unless separate text-based navigation is also provided on the same page.

(e) When specifying document colours in a BODY tag, numeric RGB notation should always be used.

(f) When using a floating image or table, "br clear" should if possible be used ahead of any further images or tables.
(g) When using HTML Tables, provision should be made to ensure the document is legible to browsers not supporting this feature.
(h) HTML containers (such as paragraphs or table cells) should be explicitly closed.
(i) Pages should not depend on a particular browser window, font size or colour table to be readable. Indeed, they should not depend on any visual presentation whatsoever, except where the information content is inherently visual in nature.
(j) Authors should not use constructs which make assumptions (explicit or otherwise) about a reader's settings. Examples to be avoided are full-screen tables or divider GIFs whose size is expressed in pixels. <TABLE WIDTH="95%"> is acceptable; <TABLE WIDTH="500"> is not.

(8) Style Sheets
(a) Authors may use style sheets to enhance web pages, and are encouraged to do so when seeking to determine document appearance.
(b) Style sheets shall not be visible to browsers which do not support them. This shall be tested.
(c) Style sheets shall not be used in a manner detrimental to accessibility for browsers not supporting this feature.

(9) Client-Side Scripting
(a) Client-side scripting languages such as Javascript may be used, provided it does not detract from the page's accessibility to browsers not supporting or enabling this feature.
(b) Script pages shall be inspected in browsers not supporting the scripting language (Not merely browsers with this facility turned off) to ensure satisfactory appearance.
(c) Script pages shall be subject to HTML validation requirements.

4. Conclusion and Future Work

In this paper, the focus is on the 'why' and 'how' factor of provenance matrix which will refine the earlier provenance matrix which consists of only four factors ie. 'when', 'where', 'who' and 'what'. This will help in more efficiently and accurately in detecting and eliminating the near-duplicates. In future, a further study will be made on the characteristics and properties of Web Provenance in near duplicates detection and elimination and also the calculation method of trustworthiness in varied web search environments and varied domains. As the future work, the architecture of a search engine can be designed or a web crawler, based on Web Provenance for the semantics based detection and elimination of near-duplicates. Also the ranking can be done based on trustworthiness values in addition to the present link structure techniques which are expected to be more effective in web search.

ACKNOWLEDGMENT

Thanks to the authorities of their working organization and management for their support and encouragement to pursue research in the chosen field of study.

REFERENCES

[1] Y. Syed Mudhasir, J. Deepika, S. Sendhilkumar, G. S. Mahalakshmi, Near-Duplicates Detection and Elimination Based on Web Provenance for Effective Web Search in International Journal on Internet and Distributed Computing Systems. Vol: 1 No: 1, 2011
[2] Salha Alzahrani and Naomie Salim, Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection, 2010.
[3] Krishnamurthy Koduvayur Viswanathan and Tim Finin, Text Based Similarity Metrics and Delta for Semantic Web Graphs, pp: 17-20, 2010.
[4] Cristina Scheau, Traian Rebedea, Costin Chiru and Stefan Trausan- Matu, Improving the Relevance of Search Engine Results by Using Semantic Information from Wikipedia, IEEE International Conference 2010, pp:151-156, 2010.
[5] George Komatsoulis, Toward a Functional Model of Data Provenance,2004.
[6] Cristina Scheau, Traian Rebedea, Costin Chiru and Stefan Trausan- Matu, Improving the Relevance of Search Engine Results by Using Semantic Information from Wikipedia, IEEE International Conference 2010, pp:151-156, 2010.
[7] Standards for HTML Authoring for the World Wide Web, copyrighted by Web Design Group (1996-2006).