IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

Application of Low Rank Online Multimodal Distance Metric Learning Technique for Semantic Hashing of Web Documents

P. Suryanka[1], L. Yamuna[2]

M.Tech Student, CSE Dept, Pragati Engineering College (Autonomous), Surampalem, A.P, India [1]

Asst. Prof., CSE Dept., Pragati Engineering College (Autonomous), Surampalem, A.P. India [2]

Abstract: Semantic hashing for web documents is essential for effective information dissemination. This paper is a sincere effort towards application of a novel method which outputs a semantic hash for an input web document. The need of such method arises as a result of research search where user may be so naïve that they are unaware of domain specific keywords or any labels for satisfying their search goals. The proposed technique in this paper assigns rank from 1 to n based on highly relevant modals of a web document. We have used six of such modals and duly considered their impact in finding semantics of a web document when hashing with a user input document. The algorithm is used in many information retrieval systems and employs a distance metric learning mechanism in practice.

Keywords: Semantic, Retrieval, Multi-Modal, Context, LOMDML.

I. INTRODUCTION

Information retrieval is crux of modern technology. Many emerging techniques in this field are recursively trying to better their performance in terms of rate at which information is accessed. Exploratory search is regarded as special field of information exploration representing group of tasks carried out by searchers who may or may not be familiar with domain of their goal, not so much sure about the ways to achieve their goals. But such users aggregate querying & browsing strategies to enhance learning as well as investigation. The main research challenges in this field are generated by questions like "what if user does not even know the domain keywords or a single answer is not what the user is looking for". As a result of this questions researchers are developing means for contextual search which are so dynamic in nature that naïve searchers can also get their desired web documents in an appreciable amount of time. Semantic search has the ability to search according to intent, contextual meaning of users understanding semantic search systems take into account various factors like location, intent, synonyms, variation of words, matching of concepts, natural language queries. In this paper, we address the issue of semantic hashing of a web document with another document using a low rank online multimodal distance metric learning algorithm (LOMDML). In our paper the algorithm will search for a user defined web document against the documents present in the web with intended highest accuracy. The documents are said to be semantically hashed if the science of meaning in language is similar for both the documents. We should observe this scenario that there is no specific document which the user knows about but still trying to get that information.

II. OVERVIEW OF PROPOSED MECHANISM

Modals considered in proposed algorithms for semantic hashing of web documents:

1. Commercial or non-commercial
2. Presence of audio or video element
3. Similar domain labels
4. Natural image or synthetic image
5. Meta description
6. Site Maps

The first modal or feature we use with a web document is its web page categorization. In broader sense all business related documents are labelled in commercial class and other nature of documents is classified under non-commercial. In commercial category, again classification can be done by areas like sports, music, arts, others.

The second modal pertains to presence of multimedia data element in a web document. It may be either an audio or video element. Verifying the contents of audio or video is out of scope of this paper. The third modal is similarity of domain labels. Each domain whether it may be science, arts, music, sports etc. Will have few explicit class labels

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

which occur in all web documents belonging to same domain. Net feature (modal) we consider is presence of image in a web document. Images can be of either nature like natural or synthetic. Natural image is a photographic.
Image captured through cameras or any other video capturing tools while synthetic image are images which are developed by users or manipulated on natural image. The next modal to be considered is the Meta description which is used to specify page description, keywords. If the Meta description of two documents is similar, we say they hash semantically. Site map is the last feature we consider for hashing two web documents in our technique. A site map essentially relays website functionally, web visitors etc. Site maps generally improve search engine optimization of a site by taking care that all pages in the specific site are found.

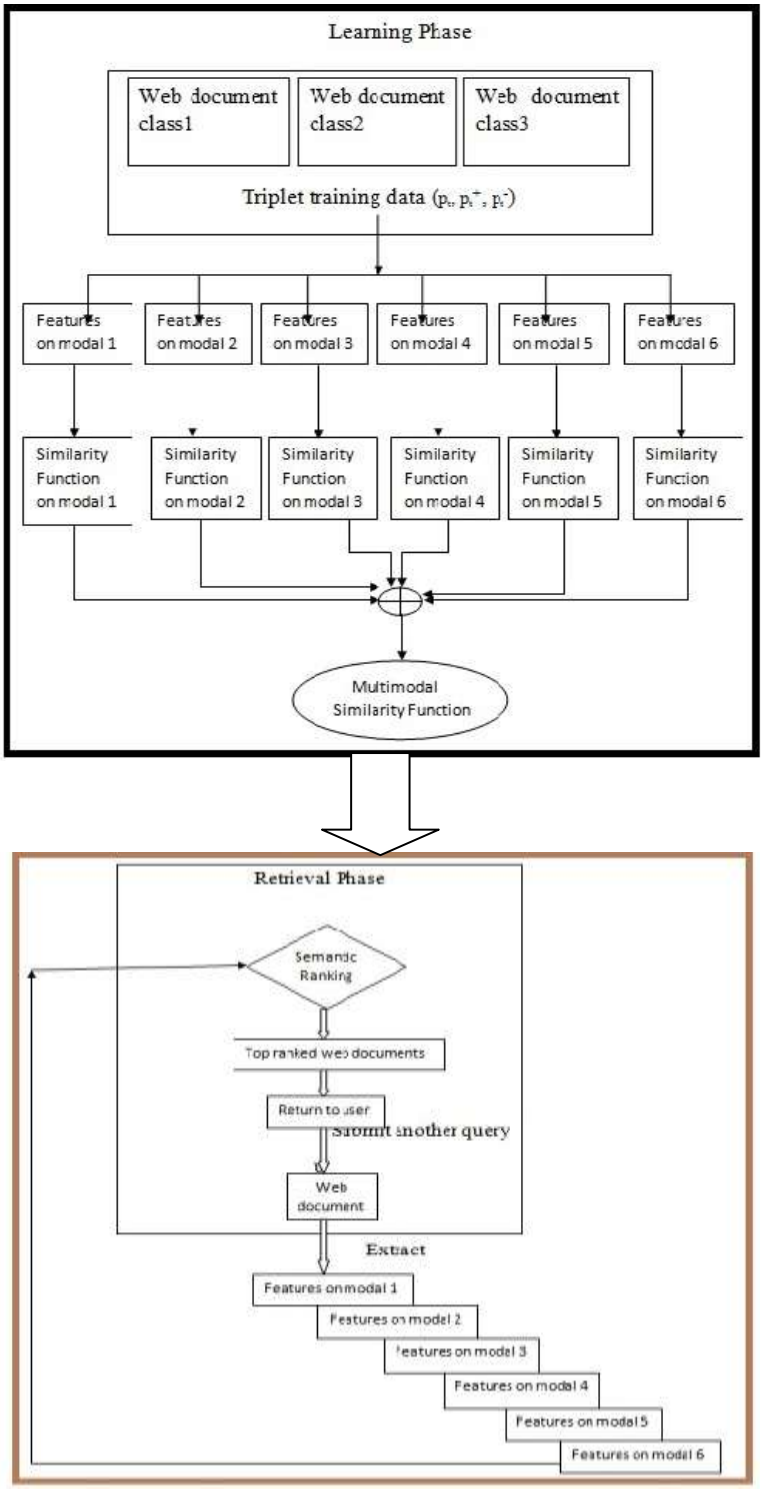

Fig.1. Overview of the Proposed Architecture

DOI10.17148/IJARCCE.2017.6514

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

Notations used in this paper:
m: no. of modalities (type of features) of the web document.
n_i : the dimensionality of the i^{th} visual feature space (modality) in a single web document.
m^i : the optimal distance metric on the i^{th} modality.
S: a positive constraint set, where a pair $(p_i, p_j) \in S$ if and only if P_i semantically hashes to p_j.
D: a negative constraint set, where a pair $(p_i, p_j) \in S$ if and only if P_i does not semantically hash to P_j.
P : a triplet set, where $P=\{(p_t, p_t^+, p_t^-)|(p_t, p_t^+) \in S; (p_t^+, p_t^-) \in D, t=1,\ldots\ldots, T$,Where T denotes cardinality of entire triplet set.
d_i (P_1, P_2): the distance function of two web documents P1 **&** P_2 on the i^{th} type of modality.

III.PROBLEM FORMULATION

Our objective is to find if the query document semantically maps to another document. We try to find the optimal distance metric M to determine distance between $\mathbf{P_1}$ **&** $\mathbf{P_2}$ as given below:
$d(\mathbf{P_1}, \mathbf{P_2}) = (\mathbf{P_1} - \mathbf{P_2})^T M (\mathbf{P_1} - \mathbf{P_2})$ where $M \geq 0$

To formulate the learning task, we consider a collection of training data objects in the form of triplets like

$$P = \{ (p_t, p_t^+, p_t^-), t=1, \ldots T \}.$$

Here, each triplet indicates hash value with respective to query web document p_t^+ indicates semantically hashing document where as $^+$, p_t^- represents semantically ill hashing document, t indicates the i^{th} document starting from 1 to last web document present in database/web database. We advocate a general principle like

$$d(p_t, p_t^+) \leq d(p_t, p_t^-) - 1 \text{ for all } t=1, \ldots. T.$$

In above equation -1 is a margin parameter to make sure a sufficiently large difference.
The final optimal distance function is represented as,

$$d(p_1, p_2) = \sum_{i=1}^{m} \theta^i \, di(P_1^{(i)} - P_2^{(i)})$$
$$= \sum_{i=1}^{m} \theta^i (P_1^{(i)} - P_2^{(i)})^T M^{(i)} (P_1^{(i)} - P_2^{(i)})$$

In above equation $\theta^{(i)} \in [0, 1]$ representing combination of weights for all six modalities.

IV.ALGORITHM

The Low- rank OMDML Algorithm:
Input:
Margin parameter - $m_p>0$
Learning rate parameter - $lr_p>0$
Discount weight parameter – $Dw_p \in (0,1)$

Step 1: Initialise $\theta_1^{(i)} = \frac{1}{m}$, $W_t^{(i)}$, for all i = 1,2,3,4,5,6
Step 2: while t=1, 2,…,6 do
Step 3: Receive: (P_t, P_t^+, P_t^-)
Step 4: Calculate $f_t^{(i)} = d_i(P_t, P_t^+) - d_i(P_t, P_t^-)$, i=1,2,…,6
Step 5: Calculate $f_t^{(i)} = \sum_{i=1}^{m} \theta_t^{(i)} f_t^{(i)}$
Step 6: if $f_t + m_p > 0$ *then*
Step 7: for i=1 to 6 do
Step 8: set $Z_t^{(i)} = II(f_t^{(i)} > 0)$
Step 9: Update $\theta_{t+1}^{(i)} \leftarrow \theta_t^{(i)}$ Dwp
Step 10: $W_{t+1}^{(i)} \leftarrow W_t^{(i)} - lr_p \nabla_t W^{(i)}$
Step 11: end for
Step 12: $\alpha_{t+1} = \sum_{i=1}^{6} \theta_{t+1}^{(i)}$
Step 13: $\theta_{t+1}^{(i)} \leftarrow \theta_{t+1}^{(i)} / \alpha_{t+1}$,i=1 to 6
Step 14: end if
Step 15: end while

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

In our stated algorithm refers to linear transformation matrix, II denotes the forbenius norm and we apply the popular Hedge algorithm for updation of the combinational weights online.
The indicator of ranking result is $Z^{(i)}_t$.
If $Z^{(i)}_t = 1$ web documents semantically hash else if its value is 0 then it does not hash semantically.

We should note here that, if $f^{(i)}_t > 0$ which means $d_i(p_t, p_t^+) > d_i(p_t, p_t^-)$, we say that the current i^{th} metric makes a error on predicting the ranking.

V. APPLICATIONS

Applications of proposed technique:
1. Improving the efficiency of QA systems.
2. Enhancing the quality of search results.
3. Implementation, expansion, robust maintenance of web directories.

VI. CONCLUSION

In this paper we have applied LOMDML algorithm to extract the semantic similarity of web documents which is key challenge of modern information retrieval systems. Our paper advocates a multi modal feature extraction of web documents in a context which hashes to similar documents present in web. Instead of formal indexing techniques available currently, we have chosen the rank based distance metric learning technique to assign ranks to the web documents which semantically correspond to each other. In future, we intend to improve the computational cost of the algorithm applied, so that searchers who intend to efficiently access information have their task cut-out. This paper will act as a readymade guide for implementers who practice hybrid mechanisms to efficiently as well as efficiently perform semantic hashing.

REFERENCES

[1] AMITAY, E. 1998. Using common hypertext links to identify the best phrasal description of target Web documents. In Proceedings of the SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web (Melbourne, Australia).
[2] AMITAY, E., CARMEL, D., DARLOW, A., LEMPEL, R., AND SOFFER, A. 2003. The connectivity sonar: Detecting site functionality by structural patterns. In Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HYPERTEXT). ACM Press, New York, NY, 38–47.
[3] BENNETT, P. N., DUMAIS, S. T., AND HORVITZ, E. 2005. The combination of text classifiers using reliability indicators. Inform. Retriev. 8, 1, 67–100.
[4] BERENDT, B. AND HANSER, C. 2007. Tags are not metadata, but "just more content"—to some people. In Proceedings of the International Conference on Weblogs and Social Media. 26–28.
[5] CHAKRABARTI, S., VAN DEN BERG, M., AND DOM, B. 1999. Focused crawling: A new approach to topic-specific Web resource discovery. In Proceeding of the 8th International Conference on World Wide Web (WWW).Elsevier, New York, NY, 1623–1640.
[6] CHEKURI, C., GOLDWASSER, M., RAGHAVAN, P., AND UPFAL, E. 1997. Web search using automated classification. In Proceedings of the Sixth International World Wide Web conference (Santa Clara, CA). Poster POS725.