International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 12, December 2016
www.ijates.com
ijates
ISSN 2348 - 7550

FACILITATING DOCUMENT RETRIEVAL BASED ON ANNOTATION AND CONTENT PHRASES

Sonal A. Nikam[1], Prof .Prof.J. V. Shinde[2]

[1]*P.G. Research Scholar, Computer Engineering,Late G. N. Sapkal COE, Nashik,*

[2]*Associate Professor, Department of Computer Engineering, Late G. N. Sapkal COE, Nashik*

ABSTRACT

In proposed work, we represent the new approach which facilitates the structured metadata by extracting documents that contains sub-sequential nearest information useful for querying the database. Traditional approaches of information extraction are quite expensive and inaccurate especially when they are only working on text without knowing the exact structure of text data. Two new techniques are proposed that facilitates the generation of structured metadata by identifying documents that are likely to contain information of user interest and this information is going to be useful for querying the database to find exact information/document. Proposed approach works on the idea that humans are more likely to add the necessary metadata during creation time, if prompted by the interface; or that it is much easier for humans (and/or algorithms) to identify the metadata when such information actually exists in the document, instead of prompting users to fill in forms with information that is not available in the document. The system works on identification of structured attributes and interesting knowledge or features that are likely to appear within the document by using 2 techniques jointly utilizing the content of the text and the query workload. Along with annotation technique, noun phrase extraction is contributed in proposed system work.

Keywords: - Document annotation, adaptive forms, and collaborative platforms

I. INTRODUCTION

There are multiple domains where user generates big data on daily basis like news feed section, blogs, social network etc. To manage unstructured data generated in the form of text or large document and provide efficient search mechanism is the basic requirement of any organization. To make efficient document retrieval user provide tags to the document same as Microsoft share point, facebook. Google base provide the attribute search based on object attribute and predefined template. Search result should generate summarized output is the basic and prime requirement of any user. To get such summarized search output, system has to maintain such documents/data in smart way. Annotation technique is one of technique that helps the user to keep document summary in annotation format and provide effective précised search result. Attribute – value pairs are generally more expressive as they can contain more information than un-typed approaches [1]. Efforts to keep such decent maintenance of such annotate documents user has to take extra efforts. To generate annotation for a document, user has to fill a form containing ten or even hundreds of fields. To fill such a huge form with single document upload is a tedious and cumbersome task. Hence annotation is most ignored pattern for document preservation and searching. There no quality assurance for arbitrary system generated annotation pattern. Because in such

system generated annotation may generate some unclear and un-useful annotation or very basic annotations by analyzing only the content of document. This effective but ignored attribute – value paired annotation scheme generate effective annotation that helps for smooth and accurate searching and maintenance. This motivated us to work on Collaborative Adaptive Data Sharing platform (CADS), which is an "annotate-as-you create" infrastructure that facilitates key value paired data annotation [1]. An important aspect of contribution is the direct use of the query workload the annotation process, along with examining the content of the document. Depending on the document and the document user involvement in annotation process, these approaches have various different perceptions on what is exactly required as an input. Nevertheless, the basic aim is to find missing tags [7] that are related with the document object. The proposed system also assume that this system is desktop based system and user annotate the documents while uploading it on server and can retrieve documents by providing search annotations from any point of the world. To organize large collection of document with annotation in effective way and to provide a mechanism to retrieve specific file in efficient way is tedious task. For document organization and effective searching or retrieval of document clustering is the solution. Noun phrases are part of speech patterns that include a noun. It includes whatever other parts of speech make sense, and can include multiple nouns. For better result retrieval in proposed system we have contributed the noun phrase extraction module. In testing phase, along with the annotations, noun phrases are also extracted which is useful during document retrieval time. This process derives the efficient document retrieval over regular search.

II. RELATED WORK

S.R. Jeffery, M.J. Franklin, and A.Y. Halevy [2] proposed a system "Pay-as-You-Go User Feedback for Data space Systems," which is a line of work towards using more expressive queries it includes annotations based data preservation. This system follows the querying strategy as "pay-as – you – go" in data management. In data spaces users provide data integration hints at querying time. But in this paper it is assumed that data sources already contain structured information and the problem is to match the query attributes with the source attribute.

K. Saleem, S. Luis, Y. Deng, S.-C. Chen, V. Hristidis, and T. Li: proposed a system [3] "Towards a Business Continuity Information Network for Rapid Disaster Recovery". In this paper they consider the Crisis Management and Disaster Recovery have gained immense importance in the wake of recent man and nature inflicted calamities. They proposed a solution or model for pre-disaster preparation and post-disaster business continuity/rapid recovery. In case of disaster need of rapid information retrieval and sharing increases This paper proposed a disaster management model which works good at some extent but it is not considering the effective retrieval.

J.M. Ponte and W.B. Croft [4] proposed a system "A Language Modeling Approach to Information Retrieval". They consider this information retrieval scenario and proposed a solution to analyze the content. They proposed a approach to retrieval based on probabilistic language modeling. Their approach to modeling was non-parametric and integrates document indexing and document retrieval into a single model. But in these making prior assumptions about the similarity of document is not warrented.

R.T. Clemen and R.L. Winkler [5] proposed a system "Unanimity and Compromise among Probability Forecasters". In this paper they work on probabilities of particular uncertain event. This helps us to find out annotation and attributes

C.D. Manning, P. Raghavan, and H. Schutze [6] proposed a solution to Laplace smoothing to avoid zero probabilities for the attributes that do not appear in the workload. It helps us to converge towards accuracy.

G. Tsoumakas and I. Vlahavas [7] propose a system "Random K-Labelsets: An Ensemble Method for Multilabel Classification". This paper proposes an ensemble method for multilabel classification. The RAndom k-labELsets (RAKEL) algorithm constructs each member of the ensemble by considering a small random subset of labels and learning a single-label classifier for the prediction of each element in the power set of this subset. In this way, the proposed algorithm aims to take into account label correlations using single-label classifiers that are applied on subtasks with manageable number of labels and adequate number of examples per label. Using this we can take into account the correlation between tags for annotations. But in this collaborative annotation is missing.

P. Heymann, D. Ramage, and H. Garcia-Molina [8] proposed a system "Social Tag Prediction". This paper gives solution for prediction of tags for particular object. We can adopt this for out suggesting annotation concept.

Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee and C.L. Giles [9] proposed a paper "Real-Time Automatic Tag Recommendation". This exactly works with the same way we want for out document annotations.

D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green [10] proposed a system "Automatic Generation of Social Tags for Music Recommendation". This paper promotes same kind of auto suggestions of tags. But this is dedicated to the musical data. We are using text based documents.

B. Sigurbjornsson and R. van Zwol [11] proposed a system "Flickr Tag Recommendation Based on Collective Knowledge". This system works for Flickr and it suggest tags for images / snapshots on flicker. It guides us for web based system structure tag recommendations.

B. Russell, A. Torralba, K. Murphy, and W. Freeman [12] proposed a system "LabelMe". A Database and Web-Based Tool for Image Annotation", It also deals web based system. This solution seeks to build a large collection of images with ground truth labels to be used for object detection and recognition research. Such data is useful for supervised learning and quantitative evaluation. This supervised learning helps us to improve suggestions for particular document.

M. Franklin, A. Halevy, and D. Maier [13] proposed a paper "From Databases to Dataspaces", A New Abstraction for Information Management". The integration model of CADS is similar to that of data spaces, where a loosely integration model is proposed for heterogeneous sources. The basic difference is that data spaces integrate existing annotations for data sources, to answer queries. Our work suggests the appropriate annotation during insertion time, and also takes into consideration the query workload to identify the most promising attributes to add.

J. Madhavan et al [14] proposed a system "Web-Scale Data Integration: You Can Only Afford to Pay as You Go". In this they propose a technique that is useful for us. In CADS, the goal is to learn what attributes to suggest. Pay-as-you go integration techniques like PayGo are useful to suggest candidate matching at query time.

III. PROBLEM FORMULATION

An efficient solution is required for user to annotate the document automatically with appropriate key value pair and save annotated document.

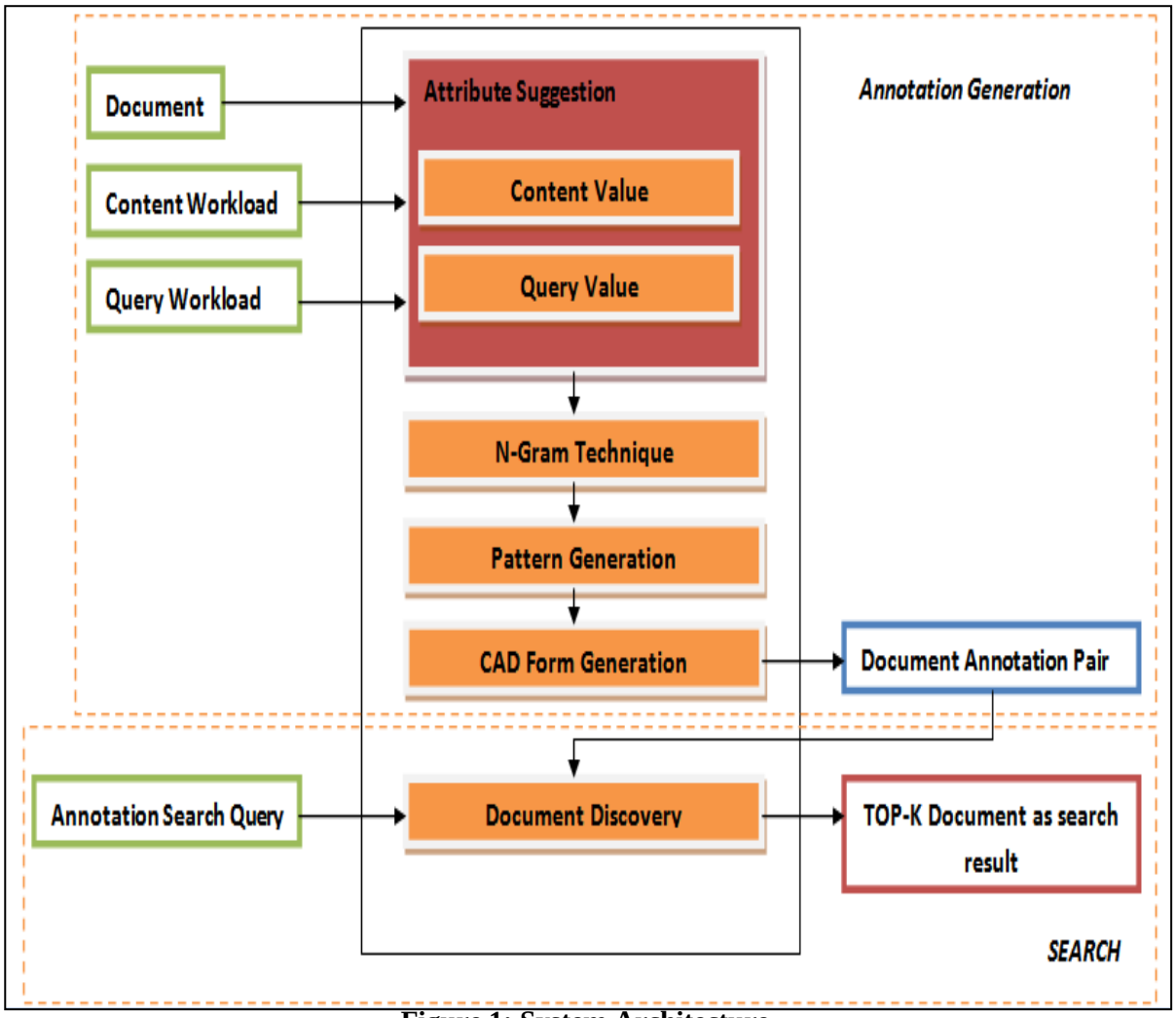

Figure 1: System Architecture

3.1 Annotation Generation

1. User first selects the document to upload it on the server. Before uploading the actual document our system analyze the document and get informative data from it.
2. To get data in annotation form in key and value pair
3. To analyze the data we first use STOP word method
4. After STOP word we use STEMMER method to filter data
5. After this we calculate the frequency count
6. Then we apply Bayes algorithm to suggest annotations from filtered data
7. After this we generate a CAD form (Collaborative Adaptive Data) which is having annotations suggested by the system. Along with the system suggestions user can add his own annotations for particular document before uploading. These annotations help us to find same document when we search it.

8. While searching, users fire some queries; these search queries are registered by our system and feed to Bernaulli Algorithm to querying value analysis. Later result of Bernaulli's algorithm is also used to suggest annotations

3.2 Search

1. User generates search query in the form of key value pair.
2. Query parameter are matched with document annotations
3. Sort match result with matching count

IV. ALGORITHMS

- **Content Value:**

The content value pd is calculated as

1. pd = p(Aj |dt) α p(Aj) · ∏ w2dt p(w|Aj)
2. Where again we assume independence among the terms.
3. We estimate p(Aj) as the smoothed frequency of Aj in the database:

$$p(Aj) = |DAj| + 1 / |D| + 1$$

- **Querying Value:**

Let W = { Q 2 W : use(Q,Aj) } be the set of queries in W that specify Aj .

The querying value Pw is calculated as:

Pw = p(Aj |W) = (|WAj | + 1) / (|W| + 1)

Combining Content Value (CV) and Querying Value (QV)

The pipelining algorithm performs sequential access on L and for each seen attribute Aj it performs a "random access" to compute CV by executing Get CV (Aj).

1) Retrieve next Aj from L

2) Get the Content Value (CV) for attribute Aj

3) Calculate the Threshold value

T = F (CV', QV (Aj)) where CV' is the maximum possible CV for the unseen attributes and QV (Aj) is the QV of Aj .

4) Let R be the set of k attributes with highest score that we have seen. Add Aj to R if possible.

5) If the k-th attribute Ak has Score (Ak) > T we return R.

Else we go back to Step 1.

Following is the pseudo code for N-Gram technique:

1] Initialize N i.e. (1 or 2 or 3)

2] Initialize variable "S" having statement

3] Initialize NGramList

4] Split the "S" and get "tokens"

5] Calculate the number of tokens for

Set K = 0

Set Cnt = token.length – N+1

Repeat

Set s = “”

Set start = K

Set end = K+N

Set J= Start

Repeat

S=S+""+tokens[j];

J = J+1

Until J < End

Set NGramList = S

K = K+1

Until : K < Cnt

6] This NGramList is mapped with the data and frequent phrase / terms is mapped with document in annotation process

Hence in search process when user fire particular term or phrase it is considered and respective document is given as search result.

V. MATHEMATICAL MODEL

System can be defined as:

S = {IU, Pr, O}

UI = { F , De , η , CAD , Q} where,

F = File / Documents

De = File Description

η = Annotations

n = noun phrase

Q = User Queries

Pr = { STE , STO , F CV , QV ,A, ,CG, Np}where

STE = Stemmer algorithm

STO = Stopwords algorithm

F = Frequency Count

CV = Content value evaluation

QV = Querying value evaluation

A = Attribute Generation

CG = CAD Generation

Np =Phrase extraction

O= {SR, NPR, CAD} where

SR: Searched result

CAD = Collaborative Adaptive Data

NPR = Noun phrase extraction results

VI. EXPERIMENTAL SETUP

System is implemented in java 1.7 with mysql as a database. System has client server architecture. Server is established using apache tomcat. A web based application is generated for user to upload document and for searching.

Dataset:

CNET [15] dataset is used for testing. It contains product reviews in database form. We have separated reviews and created multiple document file for user upload.

VII. CONCLUSION

Proposed system enhanced the efficiency and reduced the searching time. As a part of contribution noun phrase is also considered for annotation process. This is achieved using N-Gram technique. N-grams of texts are extensively used in text mining and natural language processing tasks. They are basically a set of co-occurring words within a given document or document set. Hence using unigram one can find related term and using N-gram one can find related phrase. Hence this is useful when particular user fire query related with the term pr phrase frequently occurred. These terms and phrases are extracted while annotation process is going on.

REFERENCES

[1] E.J.Ruiz, V.Hristidis, P.G.Ipeirotis, "Facilitating Document Annotation Using Content and Querying Value",IEEE transaction on knowl. And data mining, vol.26, No.2, FEB 2014.

[2] S.R. Jeffery, M.J. Franklin, and A.Y. Halevy, "Pay-as-You-Go User Feedback for Dataspace Systems," Proc. ACM SIGMOD Int'l Conf. Management Data, 2008.

[3] K. Saleem, S. Luis, Y. Deng, S.-C. Chen, V. Hristidis, and T. Li, "Towards a Business Continuity Information Network for Rapid Disaster Recovery," Proc. Int'l Conf. Digital Govt. Research (dg.o '08), 2008

[4] J.M. Ponte and W.B. Croft, "A Language Modeling Approach to Information Retrieval," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '98), pp. 275-281, http://doi.acm.org/10.1145/290941.291008, 1998

[5] R.T. Clemen and R.L. Winkler, "Unanimity and Compromise among Probability Forecasters," Management Science, vol. 36, pp. 767-779, http://portal.acm.org/citation.cfm?id=81610.81609, July 1990.

[6] C.D. Manning, P. Raghavan, and H. Schu¨tze, Introduction to Information Retrieval, first ed. Cambridge Univ. Press,http:// ww.amazon.com/exec/obidos/redirect?tag=citeulike07- 20&path=ASIN/0521865719, July 2008.

[7] G. Tsoumakas and I. Vlahavas, "Random K-Labelsets: An Ensemble Method for Multilabel Classification," Proc. 18th European Conf. Machine Learning (ECML '07), pp. 406-417, http://dx.doi.org/10.1007/978-3-540-74958-5_38, 2007

[8]] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social Tag Prediction," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 531-538, http://doi.acm.org/10.1145/1390334.1390425, 2008.

[9] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C.L. Giles, "Real-Time Automatic Tag Recommendation," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 515-522, http://doi.acm.org/10.1145/ 1390334.1390423, 2008

[10] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green, "Automatic Generation of Social Tags for Music Recommendation," Proc. Advances in Neural Information Processing Systems 20, 2008.

[11] B. Sigurbjo¨rnsson and R. van Zwol, "Flickr Tag Recommendation Based on Collective Knowledge," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 327-336, http://doi.acm.org/10.1145/ 1367497.1367542, 2008.

[12] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," Int'l J. Computer Vision, vol. 77, pp. 157-173, http://dx.doi.org/10.1007/ s11263-007-0090-8, 2008, doi: 10.1007/s11263-007-0090-8.

[13] M. Franklin, A. Halevy, and D. Maier, "From Databases to Dataspaces: A New Abstraction for Information Management," SIGMOD Record, vol. 34, pp. 27-33, http://doi.acm.org/10.1145/ 1107499.1107502, Dec. 2005.

[14] J. Madhavan et al., "Web-Scale Data Integration: You Can Only Afford to Pay as You Go," Proc. Third Biennial Conf. Innovative Data Systems Research (CIDR), 2007

[15] Dataset: http://times.cs.uiuc.edu/~wang296/Data/