International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Survey on Facilitating Document Annotation Techniques

Priyanka C. Ghegade[1], Vinod S. Wadne[2]

[1]PG Student, Department of Computer Engineering,
JSPM's Imperial College of Engineering and Research, Savitribai Phule Pune University, Wagholi , Pune , India

[2]Assistant Professor, Department of Computer Engineering
JSPM's Imperial College of Engineering and Research, Savitribai Phule Pune University, Wagholi , Pune , India

Abstract: ***Annotations can be comments, important notes, explanation about data or other types of related information or remarks that can be attached to document or to a specific part of a document. As these annotations are external, it is possible to annotate any Web document independently, without need of editing the document itself. Annotations are seen as metadata as they give additional information about an existing piece of data. As a large amount of data is produced and shared in different organization which is in text format contains the metadata about organizing product and services. This type of text structured information is getting hidden in unstructured text. Annotations are useful in effective information retrieval. Document annotation techniques add keywords or comments along with a document or part of information, this attached information is checked first for information retrieval. There are number of techniques which are useful for obtaining best annotation for documents. Techniques contain extracting information from raw data, extraction of structured metadata and many more. In this paper, we are doing a survey on document annotation techniques.***

Keywords: Annotation, Document, query, feature extraction, NLP, fuzzy logic, Structured.

1. Introduction

Nowadays the presented output on searching some type of a particular document is a primary requirement. To get such collected search output, we have to maintain documents and data in smart way i.e. stored data in structured and unstructured format. Annotation technique is one of the best featured techniques to manage such documents and get effective search result. Attribute – value pairs are generally more meaningful and significant as they can contain more information than un-typed approaches. Efforts to keep such decent maintenance of such annotate documents user has to take extra efforts.

There are many application domains like organizations and IT industries are there that generates and share information for e.g. newspapers, social networking groups like twitter face book , media channels etc. Microsoft sharing tool is one of the sharing tool that enable the user to share the information and tag or annotate it. Annotation is information related to data present and therefore it is useful in organizing the documents. Another sharing tool is Google base [1]. Google base is a database used by the Google in that user can able to add any types of data, such as text, pictures, videos, etc. It allows the users to define or suggest the attributes of data, also enable the users to select attribute values from predefined templates. But these types of tagging or annotation process requires huge amount of knowledge discovery due to the huge database information discovery.

There are many annotation techniques are present that are based on attribute value pair. The strategies based on attribute value pair are effective method of document annotation. But there is restriction that document should be in structured format when using these systems. Also user has internal knowledge of attributes of document, as there are number of attributes because of them it will be difficult and infeasible to identify such attributes and its difficult approach to facilitate document annotation. Along with these restriction it also creates more load on proposed system so that the throughput of system reduces. Even if attributes are provided, but the user has less interest in doing such things. All such difficulties will result in poor annotation. Such poor annotation results in cumbersome not only system but also data.

While annotating document special care should be taken and annotation keyword suggested should be semantic. Hence algorithm should focus on those document that contains words that are used during query. If we ignore contents of document then it will be unable to find out required information thats why document feature extraction is done on documents. In feature extraction main four things are considered proper nouns, numerical data, term weight and thematic word. Likewise for effective information retrieval and document annotation ontologies are used.

The rest of the paper is organized as follows. Section 2 discusses some related work and section 3 provides hints of some extension of proposed approaches as future work and conclusion.

2. Related Work

Eduardo J. Ruiz, Vangelis Hristidis , and Panagiotis G. Ipeirotis proposed approach in paper "Facilitating Document Annotation Using Content and Querying Value"[1] that is based on CADS (Collaborative Adaptive Data Sharing platform), which is an "annotate-as-you create" infrastructure that makes easy to present fielded type of data annotation. In the process of examining the content or data of the document, a key contribution of their system is the direct use of the type

of query workload to direct the annotation process. They were trying to prioritize the annotation of documents towards generating attribute- value pair of attributes that are often used by querying users. The primary goal of CADS infrastructure is to encourage, support and lower the cost of creating sophisticated and nicely annotated documents that can be useful for commonly issued and type of queries entered semi-structured queries. Their primary key goal is to encourage, support and provide the annotation of the documents provided or entered at creation time, though the techniques also be used for post generation document annotation while the creator of a particular document is in the phase of "document creation". Facilitation of Document Annotation using content and querying value system architecture is shown below-

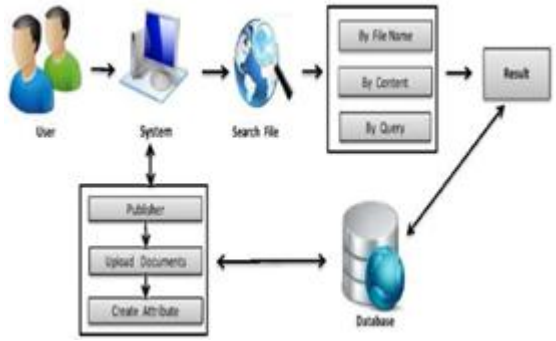

Figure 1: System Architecture of facilitating document annotation using content and querying value

In their system the author generates a new document and uploads it to the repository. After the upload, CADS analyzes that and creates an adaptive insertion form. The form contains the best type of attribute names provided to the document text and the information need i.e. query workload and the most of the probable attribute-values pair given the document text. The author i.e. creator can observed the form of information, modify or change the generated metadata as necessary and required and submit the annotated document for storage. CAD's model works as-

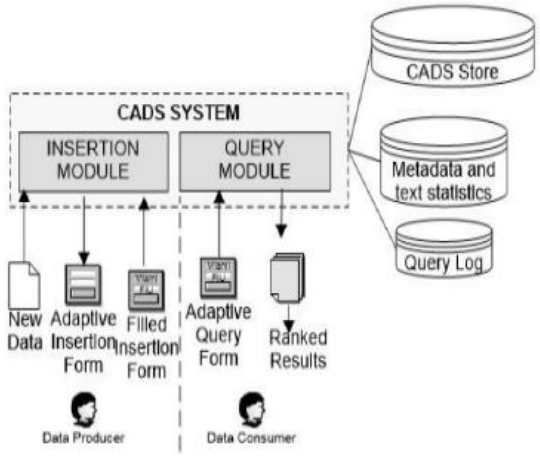

Figure 2: CADS Workflow

S.R. Jeffery, M.J. Franklin, and A.Y. Halevy proposed a paper "Pay-as-You-Go User Feedback for Data space Systems"[2] System proposes a system which is a line of work pointing towards using more expressive queries that leverage annotations is the "pay-as –you – go " querying strategy in data spaces. 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 paper "Towards a Business Continuity Information Network for Rapid Disaster Recovery"[3] Proposed paper Author take into consideration the factors Crisis Management as well as 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.

A. Jain and P.G. Ipeirotis introduced a model in "A Quality-Aware Optimizer for Information Extraction" [4] They proposed a model for estimating as well as calculating the quality of the output and retrieved results of an information extraction system when paired with a type of document retrieval strategy. They showed procedure to generate and produce a ROC curve that will helpful in generating a statistically robust and nice performance characterization of an extraction system, and then built next statistical models that use the ROC curves concept to build the *quality curves* that predict the performance of combination of an extraction system with a retrieval strategy. Our analysis helps us predict the execution time as well as output quality of an execution plan. Based on our analysis, we then show how to use these predictions to pick the fastest execution plan that generates output that satisfies the quality characteristics.

R.T. Clemen and R.L. Winkler proposed a paper "Unanimity and Compromise among Probability Forecasters "[5] In proposing approach contributions is about probabilities of particular uncertain events. This helps us to find out annotation and attributes. The paper proposes data spaces and their support systems as a new concept for data management topic. This topic contains most type of the research is going on in data management today.

P.G. Ipeirotis, F. Provost, and J. Wang experimented "Quality Management on Amazon Mechanical Turk " [6] They proposed an new algorithm for quality management of the labeling process on crowd sourced environments. The algorithm can be applied when the workers should answer a multiple choice question to complete a task. The novelty of the proposed approach is the ability to assign a single scalar score to each of the worker, which related to the quality of the assigned labels. The score did the function separates the intrinsic rate of error from the bias of the worker, allowing for more reliable quality estimation. This also leads to more fair treatment of the workers.

R. Fagin, M. Naor "Optimal Aggregation Algorithms for Middleware " [7] Paper contains simple and good algorithm TA as well as algorithms for the scenario where random access is forbidden or expensive relative to sorted access

(NRA and CA).Author introduced the instance optimality framework in the context of aggregation algorithms and provided positive as well as negative results. This proposed framework is most appropriate for analyzing and comparing the performance of algorithms and provides strong notion of optimality. We also considered approximation algorithms, and provided positive as well as negative results about instance optimality there also.2 interesting lines of investigation are: (i) finding other scenarios where instance optimality can yield meaningful results, and (ii) finding other applications of our algorithms, such as in information retrieval.

K.C.-C. Chang and S.-w. Hwang "Minimal Probing Supporting Expensive Predicates for Top-K Queries "[8] Presented framework as well as algorithms for evaluating ranked queries with expensive probe predicate. We identified that supporting probe predicates are very required and to incorporate user-defined functions, external predicates as well as fuzzy joins. Not like the existing work done which assumes only search predicates that provide sorted access to algorithm, our work addresses generally supporting expensive predicates for ranked queries. Author proposed Algorithm *MPro* which minimizes probe accesses as possible. Author developed the principle of necessary probes for determining if a probe is truly necessary in answering a *top*-query. Proposed algorithm is provably optimal, based on the necessary probe principle. Author show that *MPro* can scale well results and can be easily parallelized.

M. Franklin, D. Maier and A. Halevy proposed a system "From Databases to Data spaces:A New Abstraction for Information Management "[13].A solution is proposed to Laplace smoothing for avoiding zero probabilities for the attributes that do not appear in the workload. Proposed solutions help to converge towards accuracy. The most of the information management challenges in organizations nowadays stem from the organizations' many diverse but often interrelated data sources. Paper proposed the innovative idea of data spaces and the development of Data Space Support Platforms (DSSP). DSSPs are having the intention to free application developers from to continually again implement basic data management functionality when dealing with complicated, divisive, interrelated data sources in the same way that traditional DBMSs provide such leverage over structured relational databases. A DSSP does not assume the situation of complete control over the data in the data space. A DSSP allows the data to be managed by the participant systems but provides a new set of collected services over the aggregate of the system.

G. Tsoumakas and I. Vlahavas propose a paper "Random K-Labelsets: An Ensemble Method for Multilabel Classification"[9]. Paper proposed 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. The proposed algorithm have aim to consider label correlations using single-label classifiers that are applied on subtasks with manageable no. of labels and no. of examples per label. By Using this user can consider the correlation between tags for annotations, but here is no use of collaborative annotation.

Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C.L. Giles proposed a paper "Real-Time Automatic Tag Recommendation " [10] The proposed system exactly same works as document annotations. They proposed a learning framework for tag recommendation for scientific and web documents. We proposed a Poisson mixture model for efficient document classification. Author proposed a novel and efficient node ranking method as well as several new metrics for evaluating the performance of their framework. The proposed system framework executes its potential in evaluations on two real-world tagging data sets, indicating its capability of handling large-scale data sets in real-time. The proposed method can recommend tags in one second on average. The relationship among documents, words, and tags can then be represented by two bipartite graphs as shown in Figure:

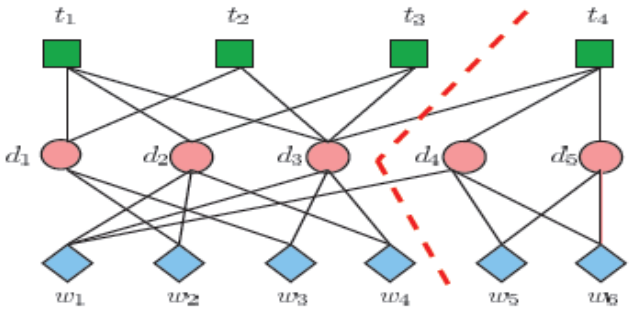

Figure 3: Two bipartite graphs of documents, words and tags.

J.M. Ponte and W.B. Croft proposed a paper "A Language Modeling Approach to Information Retrieval ". Where author takes into consideration this information retrieval scenario and proposed a solution to analyze the content. They proposed an approach to retrieval based on probabilistic language modeling. The authors approach to modeling was non-parametric and integrates document indexing and document retrieval into a single model. But in thus making prior assumption about the similarity of document is not warranted.

D. Eck, P. Lamere, T. Bertin-Mahieux proposed a paper "Automatic Generation of Social Tags for Music Recommendation" [11] The proposed paper suggests the same kind of auto suggestions of tags. This is dedicated to the musical data. We are using text based documents. The type of work proposed is preliminary, but the user believes that a supervised learning approach to auto tagging has substantial merit like other system. The next step of the system is to compare the performance of our boosted model to another type of approaches such as SVMs and neural networks. The data set used for these experiments is already larger than those used for publishing results for genre and artist classification. A dataset another order of magnitude larger is necessary to approximate even a small commercial database of music. Next further step of the system is comparing the performance of our audio features with other sets of audio features.

B. Sigurbjornsson and R. van Zwol proposed a paper "Flickr Tag Recommendation Based on Collective Knowledge" [12] Proposed system works for Flickr and it suggests tags for images / snapshots on Flickr. It guides us for web based system structure tag recommendations. Annotating photos through tagging is a popular way to index and organize photos. Author presented a characterization of tag behavior in Flickr, which forms the foundation for the tag recommendation system and evaluation presented in the second part of the paper.

A. Jain and P.G. Ipeirotis propose a paper "A Quality- Aware Optimizer for Information Extraction"[4].The respective paper presents and explains the Receiver Operating Characteristic (ROC) curves to calculate the extraction quality and selection of extraction parameters. Automated information extraction (IE) algorithms used to extract targeted relations or characteristic of the document. In this case we should process only documents that actually contain such information. When we process documents that do not match with the predefined targets.

Another document annotation technique based on extraction of document features. Features are consider term weight, proper noun, numerical data and thematic word. Before extracting features preprocessing is done. In preprocessing module, all unwanted things like stop words, special symbols are removed and stemming is done. After features of document are extracted and finds feature score and same is provided as input to fuzzy logic. From this section we get summary. By using summary and annotation label document is an noted.

3. Conclusion

This paper surveys techniques of document annotation. We have studied document annotation strategies based on attribute-value pair and document features. Those are useful for annotating document at uploading time as well as considers the things requires for users querying. Finally conclude that the proposed document annotation techniques is efficient and useful in effective information retrieval.

4. Acknowledgment

I express my gratitude towards Prof. V. S. Wadne , project guide and Prof. R. N. Phursule , P.G. coordinator and Prof. S. R. Todmal, Head of Department of Computer Engineering , Imperial College Of Engineering And Research Wagholi, Pune who guided and encouraged me in completing this paper. I would like to thanks our Principal Dr. S. V.Admane for allowing us to publish paper.

References

[1] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis proposed "Facilitating Document Annotation Using Content and Querying Value", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 2, FEBRUARY 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] A. Jain and P.G. Ipeirotis, "A Quality-Aware Optimizer for Information Extraction," ACM Trans. Database Systems, vol. 34, article 5, 2009.

[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] P.G. Ipeirotis, F. Provost, and J. Wang, "Quality Management on Amazon Mechanical Turk," Proc. ACM SIGKDD Workshop Human Computation (HCOMP '10), pp. 64-67, http://doi.acm.org/ 10.1145/1837885.1837906, 2010.

[7] R. Fagin, A. Lotem, and M. Naor, "Optimal Aggregation Algorithms for Middleware," J. Computer Systems Sciences, vol. 66, pp. 614-656, http://portal.acm.org/citation. cfm?id= 861182.861185, June 2003.

[8] K.C.-C. Chang and S.-w. Hwang, "Minimal Probing: Supporting Expensive Predicates for Top-K Queries," Proc. ACM SIGMOD Int'l Conf. Management Data, 2002.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

Author Profile

Mr. Vinod S. Wadne received the B.E and M.E. degrees in computer Engineering. And working As Assistant professor In Department of Computer Engineering at Imperial College Of Engineering And Research,

Wagholi, Pune. He has 11 year teaching Experience. His areas of interest in research is Data Mining.

Priyanka C. Ghegade received the B.E. degree in computer Engineering from Pune university and pursuing master of Engineering in computer Engineering from Imperial College Of Engineering And Research, Wagholi, Pune and Working as Assistant professor In Department of Computer Engineering at HSBPVT's COE college. She has 9 Months of experience. Her areas of interest in research is Data Mining.