A Survey on Facilitating Document Annotation Using Content and Querying Value Based on Attributes Suggestion Strategy

Shital P. Dhok*, Mrudula Nimbarte
Department of Computer Science & Engineering
Bapurao Deshmukh College of Engineering
Sevagram, Wardha, Maharashtra, India.
*mudganti.sheetal@gmail.com

Abstract

Now days many organizations generate and share descriptive and textual data of their products, services, and actions. Such type of collections of textual data contain significant amount of structured information, which remains saved in the unstructured text. While information extraction algorithms facilitate the extraction of structured relations in an very expensive and inaccurate way especially when operating on top of text that does not contain any instances of the targeted structured information. There are many alternative approaches that facilitate the generation of the structured metadata by identifying documents that are likely to contain information of interest and this information is going to be subsequently useful for querying the database which relies on the idea that humans are more likely to add the necessary metadata during creation time. This can be done like prompting by the interface; or that it should made much easier for humans (and/or algorithms) to identify the metadata when such information actually exists in the document, instead of naively prompting users to fill in forms with information that is not available in the document. There are different algorithms that identify structured attributes that are likely to appear within the document, by jointly utilizing the content of the text and the query workload.

Keywords-Document annotation, Adaptive forms, Collaborative Adaptive Data Sharing platform

Introduction

Summarized output on searching particular document is prime requirement nowadays. To get such summarized search output, we have to maintain documents or data in smart way. 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.

A scenario is cumbersome, complicated and tedious where there are number of fields to be filled at time of uploading a particular document. Hence end user frequently ignores such annotation capabilities. User is still unresponsive and ignoring task though system offers the facility to randomly annotate the data with attribute-value pairs. Along with this there it also has unclear usefulness for subsequent searches in the future. Such difficulties finally tend to very basic annotations, if any at all, that are often limited to simple keywords. Such simple annotations make the analysis and querying of the data cumbersome. It's the fact that this effective but ignored attribute – value paired annotation scheme can bring smooth searching and maintenance and this motivated us to work on Collaborative Adaptive Data Sharing platform (CADS), which is an “annotate- as-you create” infrastructure that facilitates fielded data annotation.

The contribution of our system is the direct use of the query workload to direct the annotation process, in addition to checking the content of the document. Along with this contribution we are also working on

phrase extraction process to build knowledge out of text. CAD provides cost effective and good solution to help efficient search result. The goal of CADS is to support a process that creates nicely annotated documents that can be immediately useful for commonly issued semi-structured queries of end user.

Related Work

A recent work (Jeffery *et al.*, 2008) proposes Pay-as-You-Go User Feedback for Data-space Systems. The system which is a line of work 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 it is assumed that data sources already contain structured information and the problem is to match the query attributes with the source attribute.

Towards a Business Continuity Information Network for Rapid Disaster Recovery. 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. They propose a disaster management model which works well at some extent but it is not considering the effective retrieval (Saleem *et al.*, 2008).

From Databases to Data spaces (Franklin *et al.*, 2005): A New Abstraction for Information Management. "It proposes 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.

A paper (Tsoumakas *et al.*, 2007) for Random K-Label sets: They give an ensemble method for multi label classification. The RAndom k-LABEELsets (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, they 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. Social Tag Prediction (Heymann *et al.*, 2008) gives a solution for prediction of tags for particular object. We can adopt this for out suggesting annotation concept. Real-Time Automatic Tag Recommendation (Song *et al.*, 2008) works with the same way we want for out document annotations.

A Language Modeling Approach (Ponte *et al.*, 1998) to Information Retrieval considers this information retrieval scenario and proposed a solution to analyze the content. They propose 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 this making prior assumption about the similarity of document is not warranted.

Automatic Generation of Social Tags for Music Recommendation (Eck *et al.*, 2008) promotes same kind of auto suggestions of tags. But this is dedicated to the musical data. We are using text based documents. A paper (Sigurbjornsson *et al.*, 2008) on Flickr Tag Recommendation Based on Collective Knowledge suggests tags for images / snapshots on flicker. It guides us for web based system structure tag recommendations.

A Quality-Aware Optimizer for Information Extraction (Jain *et al.*, 2009) presents Receiver Operating Characteristic (ROC) curves to calculate the extraction quality and selection of extraction parameter. 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 matched with the predefined targeted information and we use automated information extraction algorithms to extract such annotation. We often face a significant number of wrong positives results, which may lead to significant quality problem in the data annotation.

A Probabilistic Model for Personalized Tag Prediction (Nikam *et al.*, 2014) suggests social tagging by incremental process. It proposes Probabilistic models. A Probabilistic tag recommendation system is introduced. It uses Bayesian approach. It only focuses on content and not the query workload that reflects the user interest.

A Database and Web-Based Tool for Image Annotation (Russell *et al.*, 2011) a tag prediction for images is proposed. Web-based tool for easy image annotation and instant sharing of annotations detects the objects and finds similarity with existing dataset. It helps for image search in web.

Usher: Improving Data Quality with Dynamic Forms (Yin *et al.*, 2010) focuses on system for form design, data entry and data quality assurance. Using existing data set of form, USHER derives a probabilistic model using the questions of the form. It is closely related to CAD form in our system. Using Usher we can identify dependencies across attributes.

Automated Creation of a Forms-Based Database Query Interface (Chen *et al.*, 2010) Expressive Query Specification through For Customization is a technique to extract query forms form existing queries in a dataset that are fires on database using 'querability' of column. Microsoft Sharepoint (Microsoft, 2008) and SAP Net weaver (Jayapandian, 2011) allows user to share documents, annotate them, and perform simple keywords queries. Hard-coded attributes can be added to specialized insertion forms.

Standing out in a Crowd: Selecting Attributes for Maximum Visibility (Miah *et al.*, 2008) propose extract algorithm based on Integer Programming formulation of the problem. It takes significant amount of time for processing for small workload but provide optimal and nearest solution.

Proposed System

CAD's basic objective is to create very structured annotated document to trigger efficient search in minimal execution cost. Also for semi-structured queries of user CAD generate most useful output. Also CAD adopt the strategy in which document is annotate at time of creation while crater is still in "document generation" phase, even though the techniques can also be used for post-generation document annotation.

In our scenario, the author generates a new document and uploads it to the repository. After the upload, CADS analyzes the text and creates an adaptive insertion form. The form contains the best attribute names given the document text and the information need (query workload), and the most probable attribute values given the document text. The author (creator) can inspect the form, modify the generated metadata as necessary, and submit the annotated document for storage. Our efforts focus not only on identifying the potential annotations fields that exist in complete and optimal annotations for document , but also to rank them and display on top the most important ones. Since the goal of annotations is to facilitate future querying, we want the annotation effort to focus on generating annotations useful for the queries in the query workload.

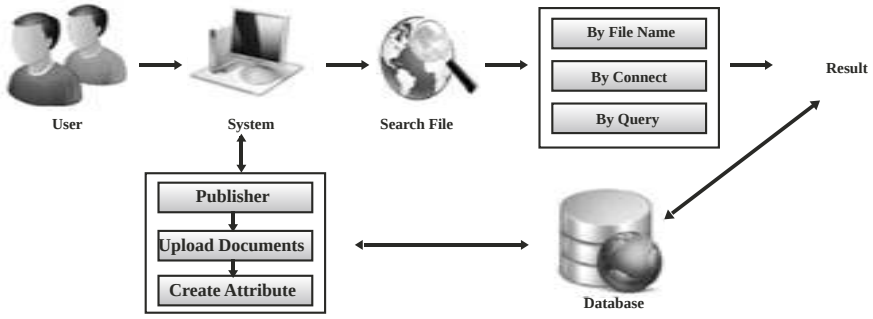

Figure 1: Proposed System

Conclusion

In this survey, we studied adaptive techniques to suggest relevant attributes to annotate a document, while trying to satisfy the user querying needs. Many papers are based on a probabilistic framework that considers the evidence in the document content and the query workload. There present two ways to combine these two pieces of evidence, content value and querying value: a model that considers both components conditionally independent and a linear weighted model. Experiments show that using these techniques, attributes that can improve the visibility of the documents with respect to the query workload by up to 50%. That is, we conclude that using the query workload can greatly improve the annotation process and increase the utility of shared data.

References

Eduardo, J, Ruiz., Vagelis, Hristidis., Panagiotis, G, Ipeirotis.2014. Facilitating Documents Annotation Using Content and Querying Value. *IEEE transaction on Knowledge and Data Engineering,* 26, 2.

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

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

Jain, A., Ipeirotis, P.G., 2009. A Quality-Aware Optimizer for Information Extraction, *ACM Trans. Database Systems,* 34, 5.

Ponte, J.M., Croft, W.B.1998. A Language Modeling Approach to Information Retrieval. Proc.21st Ann.Int' l ACMSIGIR Conf. *Research and Development in Information Retrival (SIGIR'98),* 27281.

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

Miah, M., Das, G., Hristidis, V., Mannila, H. 2008. Standing out in a Crowd: Selecting Attributes for Maximum Visibility. Proc. Int'l Conf. Data Eng. (ICDE).

Heymann, P., Ramage, D., Garcia-Molina, H. 2008.Social Tag Prediction," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval(SIGIR'08),5315-538.

Song, Y., Zhuang, Z., Li, H., Zhao, Q., Li, J, Lee, W.-C., Giles, C. L. 2008. Real-Time Automatic Tag Recommendation. Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), 515-522.

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

Sigurbjornsson, B., van Zwol, R. 2008. Flicker Tag Recommendation Based on Collective Knowledge," Proc. 17th Int'l Conf. World Wide Web (WWW '08), 327-336.

Russell, B., Torralba, A., Murphy, K., Freeman, W. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int'l J. Computer Vision,* 77, 157-173.

Franklin, M., Halevy A., Maier, D. 2005. From Databases to Dataspaces: A New Abstraction for Information Management, SIGMOD Record, 3, 2734.

Microsoft, Microsoft SharePoint, http://www.microsoft.com/sharepoint/, 2012.SAP, Sap Content Manager, https://www.sdn.sap.com/irj/sdn/nw-cm, 2011.

Jayapandian, M., H.V, Jagadish. Automated Creation of a Forms-Based Database Query Interface. Proc.VLDB Endowment, ol.1, 695-709.

Chen, K., Chen, H., Conway, N., Hellerstein, J.M., Parikh, T.S. 2010. Usher: Improving Data Quality with Dynamic Forms. Proc. IEEE 26th Int'l Conf. Data Eng. (ICDE).

Yin, D., Xue, Z., Hong, L, Davison, B.D. 2010. A Probabilistic Model for Personalized Tag Prediction. Proc.ACMSIGKDD Int'l Conference Knowledge Discovery Data Mining.

Nikam S, Prof. Shinde JV. 2014. Survey on Facilitating Document Annotation Using Content and Querying Value. *International Journal Computer Science Information Technology.* 5, 8150-8152.

Vishal, A., Pankaj, K, Patil. 2015. Survey on Facilitating Document Annotation Using Content and Querying Value. *International Journal of Science and Research,* 4.