© 2015 IJSRSET | Volume 1 | Issue 1| Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Attribute Based Document De-Duplication Using the Metadata based Framework

Ravikanth M[1], Bhuvaneshwari[2]

[1]Associate Professor of CSE in CMRTC, Hyderabad, Telangana, India

[2]Professor of CSE in CU, Kalapet, Pondicherry, Tamil Nadu, India

ABSTRACT

Technology and its advantage of using in the modern age of Information Technology, where the content based document de-duplication keep on changing. In the context of the structured and unstructured data gives us the most significant information, but in order to process the data of the content structured would be useful. In this Paper, we try to give the most significant glimpse of the metadata based information in the Human Interface of the UI. Technologically its process of facilitation but cannot ensure all mentioning your data can be made search. In order to over to such trend we need protocol of User interface before submitting the data making in the format the query based structured or unstructured approach. In this one we have used the UI based framework which in turn uses the approach of the content in the document in order to facilitate the process of the QTP and the metadata makes the sense protocol of the category.

Keywords: Document, de-duplication, adaptive forms, collaborative platforms

I. INTRODUCTION

The challenge of automatically annotating documents with structured semantic metadata has been addressed in previous projects by BBC Research and Development, including work on automated concept tagging and document linking. However, to date, this work has consisted of using supervised learning models which require a training set which must be compiled by hand in advance, or substituted by some heuristic or external source of information which serves as a suitable proxy for ground truth A thorough, empirical analysis of the effectiveness of active learning techniques already published in the literature, including a comprehensive appraisal of different query strategies and error measures, as well as a comparison to established supervised and semi-supervised learning algorithms. By combining active learning approaches to inference with models that reflect the structure of the topic space in question, we hope to make a novel contribution to the field of automated topic modeling and classification. Specific approaches could include the use of query-by-committee to estimate classification variance in a model where computing such statistics directly would be intractable, combining active learning with complementary semi-supervised and unsupervised learning techniques, and the use of active learning to infer a topic distribution over documents, by defining a topic model over users of the system and propagating this to documents via their training decisions.

II. RELATED WORK

In addition to different goals and different prediction aims, two more criteria must be considered in the design phase of an adaptive component. First, in some cases, very detailed information about the users of the system is available { for instance, in cases where the same course has been conducted several times and analyses of usage data have shown that users' behavior changes only marginally across various semesters, or in cases where users have already taken several different courses on the platform, whereas in other cases the system does not know anything about its users.

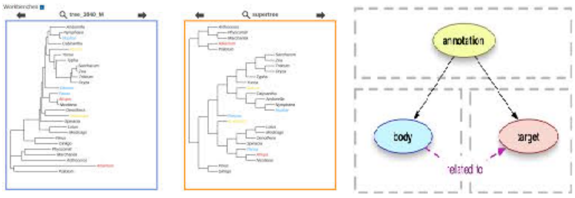

Figure 1. Related view of the Document De-duplication

This cans crucially influence the decision on how usage data is analyzed and evaluated. The modeling step aggregates usage data by bundling data instances based on their relations to each other. The outcome is a new data instance which can then be passed over to the analysis unit responsible for prediction. The modeling approach again depends on the nature of the data and the prediction aims. If usage data comprises information of different tools and users, it might be interesting to find out how users are related to each other, or how tools are used in combination. If usage data comprises data of only one tool that can be used in different ways, it might be interesting to find out in what order the different activities happened.The combination of students and problems can help to not only find out which approach a student shows, but also if the approach deferrers for different types of problems. Both entropy indices, however, tend to naturally increase as the number of clusters increases in the concrete scenario, and are, therefore, not sufficient in themselves for characterizing the results of the clustering process. This increase originates, in this case, from clusters often being homogeneous along one dimension but inhomogeneous along others. For instance, regarding the distribution of students to clusters, we have to consider that problem-solving behavior consists of several components that could influence the assignment of a student's problem-solving sequence to a cluster. The more clusters are introduced; the granularity of the analysis, and the more factors could end up being emphasized by the representation in a cluster. Thus, it is practically impossible to receive the optimal cluster setting mentioned before. However, a good cluster setting would be able to group a student's problem-solving sequences into the same cluster if they are similar at least along one dimension.

III. METHODOLOGY

In the Era of technology, the individual cluster setting evaluation metrics were introduced: the student entropies, the problem entropies, the variance, and the expected prediction error. The entropy-based metrics aim at capturing intra-personal similarities and the effectss of problem types on the behavior of learners. Regarding student entropies, patterns are indicative of students showing stable problem-solving behavior. Regarding problem entropies, recognized patterns “are either indicative of problems of the same structure or of independent approaches people share. The expected prediction error can be assessed in a way similar to what was described for variance. Generally, a low value for the expected prediction error is good. However, if the number of clusters becomes too high, thus minimizing the number of data instances in a cluster and thus also the expected prediction error, the resulting clusters are not informative in any way anymore. Thus, again, the value for the metric should not be a certain threshold. it should be assured that adaptations based on predictions are in general potentially reasonable. For instance, if the system recognizes a pattern in the current live graph based on the users involved, how would a subsequent adaptation look like? For example, to recommend collaborators for future activities, an ideal system would strive to take into consideration any observable extrinsic factors that may a effect collaboration. Summing up, the described approach could potentially provide valuable results regarding predictions and subsequent adaptations if specific conditions are held. If this is not the case, the approach's success can only be measured after practical implementation.

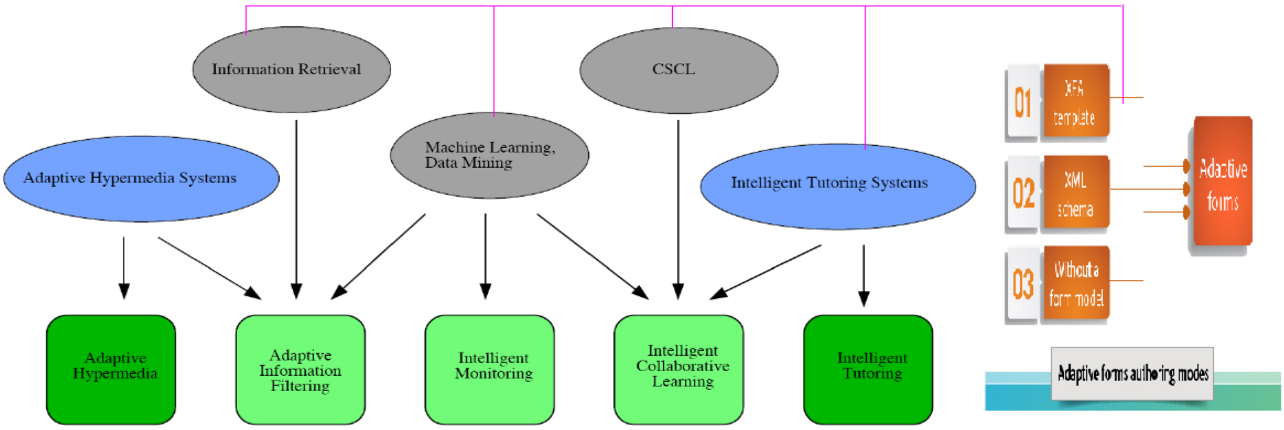
Information Retrieval
CSCL
Machine Learning, Data Mining
Adaptive Hypermedia Systems
Intelligent Tutoring Systems
Adaptive Hypermedia
Adaptive Information Filtering
Intelligent Monitoring
Intelligent Collaborative Learning
Intelligent Tutoring
01 XFA template
02 XML schema
03 Without a form model
Adaptive forms
Adaptive forms authoring modes

Figure 2. Model Retrieval de-duplication Metadata System

related, criteria to determine the probabilities for relations must be defined. These criteria can be different for different tools, e.g. for an asynchronous communication tool, the time frame in which related activities can occur is much longer than in synchronous communication tools. Thus, we find a default setting for splitting tool activities into time slots.As classification should be performed on the basis of interrelated activity sequences, the activities cannot be sent through the classification process. Another approach would be to include several items' data as features in a data set which then goes to the classifiers. However, as the number of items in every time slot can different drastically, classification on the basis of time-slot-data is not possible. Yet, a fixed number of items can be used for one data set and include the corresponding time slot index as an additional feature in the feature set for classification data. Another possibility is to build the feature set based on the metrics described before, i.e. activity data will not be sent through the classifier in its original form, but be preprocessed first, creating a new data set for e.g. every time slot, containing elements like mean, standard deviation, variance, minimum and maximum, number of elements in this time slot, etc. Furthermore, standard matrix and graph metrics like scarcity or connectivity can be included. Given the results of the classifiers, one can extract information about what leads to successful collaboration and what does not seem to influence the success of a group at all. This information can then be used to add contextual information to the knowledge gained by classification of independent activities. Furthermore, it could potentially reveal new information about collaboration in general and collaborative learning in specific.

3.1 Evaluation and Analysis

The expected prediction error can be assessed in a way similar to what was described for variance. Generally, a low value for the expected prediction error is good.

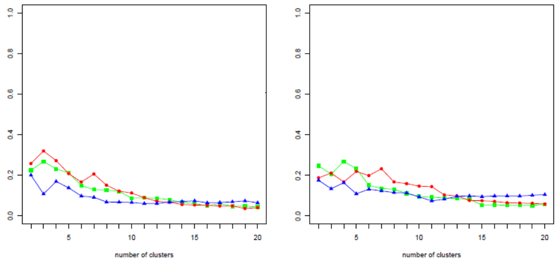

Figure 3. Comparison of the cluster based metadata

However, if the number of clusters becomes too high, thus minimizing the number of data instances in a cluster and thus also the expected prediction error, the resulting clusters are not informative in any way anymore. Thus, again, the value for the metric should not decree a certain threshold.

IV. CONCLUSION AND FUTURE WORK

Several alternatives to representing sequences have been considered and evaluated, concentrating on the interwoven questions of comparability. Specifically, a formalism should be found, that would allow for the comparison of activity sequences that might differ only little, but also for comparing sequences with only small amounts of overlap. In general, the modeling of sequential data faces the challenge of not losing information about relations and dependencies between the individual items

V. REFERENCES

[1]. "Google," Google Base, http://www.google.com/base, 2011.

[2]. S.R. Jeffery, M.J. Franklin, and A.Y. Halevy, "Pay-as-You-Go UserFeedback 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]. 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.

[6]. R.T. Clemen and R.L. Winkler, "Unanimity and Compromise among Probability Forecasters," Management Science, vol. 36, pp. 767-779, July 1990.

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

[8]. 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, 10.1145/1837885.1837906, 2010.

[9]. 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.

[10]. 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.

[11]. 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/10.1007/978-3-540-74958-5fi38, 2007.

[12]. M. Miah, G. Das, V. Hristidis, and H. Mannila, "Standing out in aCrowd: Selecting Attributes for Maximum Visibility," Proc. Int'l Conf. Data Eng. (ICDE), 2008.

[13]. 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.

[14]. 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.

[15]. 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.

[16]. 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.

[17]. B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A Database and Web-Based Tool for Image De-duplication," 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.