# Enriching Topic Modelling with Users' histories for Improving Tag Prediction in Q&A Systems

Glenn Boudaer, Johan Loeckx
Artificial Intelligence Lab
Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium
jloeckx@ai.vub.ac.be

## ABSTRACT

The automatic attribution of tags in Question & Answering (Q&A) systems like StackExchange can signi cantly reduce the human e ort in tagging as well as improve the consistency among users. Existing approaches typically either rely on Natural Language Processing solely or employ collaborative ltering techniques. In this paper, we attempt to combine the best of both worlds by investigating whether incorporating a personal pro le, consisting of a user's history or its social network can signi cantly improve the predictions of state-of-the-art text-based methods. Our research has found that enriching content-based text features with this personal pro le allows to trade-o the precision of predictions for recall and as such improve the "exact match" (predicting the number of tags and the tags themselves correctly) in a multi-label setting from a baseline of 18.2% text-only to 54.3%.

## 1. INTRODUCTION

Folksonomies | classi cation systems where classes and categories are not xed but eligible to evolve rapidly over time due to contributions of the community | have become a common way in Web 2.0 applications to improve the organisation of an ever increasing amount of unstructured data online. It is di erent from traditional classi cation systems like the Dewey Decimal Classi cation in the sense that so-called "tags" are attributed by all users freely, rather than in a formalized manner by a central authority. A well-known example is \Stack Exchange", a social platform centred around *collective intelligence* on which users can ask questions in a particular domain of expertise and assign appropriate tags in order for the question to be discovered by other users to answer. Though individual persons are free to choose the tags connected to their post, an interesting phenomenon is the fact that most users generally come to a consensus regarding the used tags.

From this perspective, the task of automatically assigning tags to entities in Q&A systems is situated between document classi cation and collaborative tagging. Entities and tags are connected through a network of users answering each others questions, a typical domain for collaborative ltering. The number of distinct tags employed, however, is much smaller in these Q&A sites than in \traditional" collaborative tagging systems like Flickr or Del.icio.us [3].

Traditionally, automatic tagging systems employ either purely content-based methods [9, 15, 10, 12, 13] or collaborative methods [14, 8, 16, 7]. In this paper, the goal is to investigate what the impact is of integrating a user's history to improve purely text-based methods. We will approach the automatic tagging task in Q&A systems as a *multi-label document classi cation problem*, in which documents are enriched with the user's personal tag history. This method should allow to combine the strengths of powerful content-based NLP methods with the knowledge of the crowd.

The structure of this paper is as follows. First, the basic workings of content-based automated tagging are reviewed. Next, an architecture is proposed on how to integrate social aspects into content-based methods. The dataset is discussed and the main ndings presented to wrap up with conclusions in section 5.

## 2. BACKGROUND

The history of document classi cation, commonly used to assign tags to a document, dates back to the seventies [5]. One of the fundamental problems encountered in this task is the intrinsic ambiguity in language. A word like \bank", for example, can be interpreted as a " nancial institution" as well as \sloping side of a river". Taking into account the context of words is therefore crucial to predict topics of a document and several techniques to represent documents have therefore been proposed, like Bag-of-Words (BoW) and n-gram models. Though these methods have proven their capabilities, these techniques often rely on a clear and well-formulated corpus in order to work.

### 2.1 Topic modelling

Identifying the underlying concepts referred to by words rather than the words themselves, has been found to better model these ambiguities. In 1998 a Topic Detection and Tracking (TDT) pilot study was performed, under the supervision of the Defense Advanced Research Projects Agency (DARPA) and the NIST, to look for possibilities to detect and track events in a stream of broadcast news stories [2]. They concluded that the topic segmentation task was tractable using known technologies in Machine Learning, Natural Language Processing and Information Retrieval. A new domain called \topic modelling" came into existence.


*WWW'16 Companion,* April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890566 .

One of the first approaches to model underlying topics rather than the words themselves, was Latent Semantic Analysis (LSA), initially proposed in 1988 by Dumais et al. [6]. Topics are derived from applying Singular Value Decomposition (SVD) on the term-document matrix that indexes the frequency (or TF-IDF) of words in each document. The problem of synonyms and polysemes (words that bare different meanings) are believed to be eliminated by LSA.

## 2.2 Labelled-Latent Dirichlet Allocation

A more modern approach is Latent Dirichlet Allocation, introduced in 2003 [4], consisting of a generative model in which every document is represented as a mix of different *topics*, each with their own distribution of words in which both topic and word distributions follow a Dirichlet prior. The distributions are learnt through Bayesian inference and therefore well adapted to small datasets as over fitting can be avoided. Another interesting property of LDA is the fact that it accounts for word disambiguation. A word can occur in different topics, depending on the context, effectively reflecting the different meanings. *Labelled-LDA* (L-LDA) is an extension of LDA that operates in a supervised setting that attempts to match topic discovery with given tag labels. It was successfully applied in the past on Twitter for profile classification outperforming other methods when training data was limited [11]. In the remainder of our discourse, L-LDA is chosen as a baseline for content-based tagging as it is a very well known method that is commonly applied method for automated tagging. In this paper we will analyse the impact of modelling the correlation between tags and of incorporating personal tag histories in the tagging process.

# 3. ARCHITECTURE

The architecture of the automatic tagging system with the proposed adaptations is shown in Fig. 1 and consists of two main parts: (a) an individual supervised L-LDA component, trained to make predictions for each tag and (b) a traditional classifier that combines all these inputs to yield the final multi-class output. As this classifier takes all individual tag predictions as input, it has the capacity to model the relations between tags. Three topologies were compared. The baseline architecture consists of a Binary Relevance multi-label classification based on single-label Labelled-LDA classifiers, one for each tag. This topology is denoted with \LLDA". The number of topics in the model was set equal to the number of tags. Next, a classifier is added that combines all these contributions in order to model correlations between tags (LLDA-TC). Finally, the impact of a user's personal tag profile is investigated (LLDA-TC-PTP). The same multi-label classifier that was used for finding patterns between the individual L-LDA predictions (in the center), is used to incorporate information from the user profile information (bottom). Different kinds of classifiers types were compared (decision tree, Naive Bayes and Support Vector Machines). It was found that Support Vector Machines delivered the best performance.

## 3.1 Tag correlation

In a multi-label classification task (when documents are labelled with more than one tag), tags are often correlated and therefore they can help disambiguation of words and concepts as well. For example, the probability of a tag

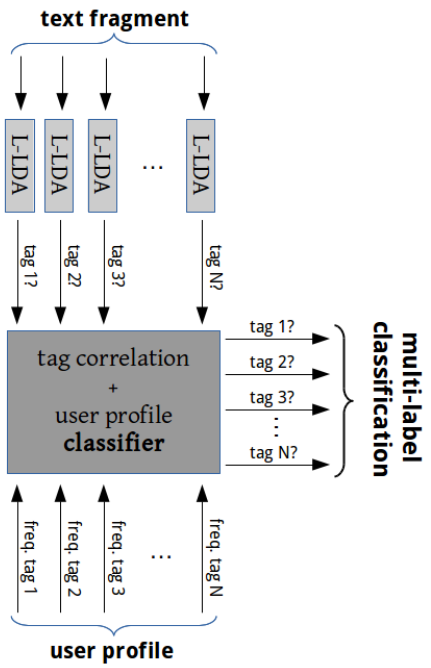


**Figure 1: The architecture of our methodology. Single-label Labelled-LDA is combined with a personal user profile consisting of its tag history to obtain the final multi-class output. The central classifier models tag correlation and impact of the user profile.**

\bank" meaning financial institution is higher if a document has also been tagged with \money" rather than with \nature". For this reason, a single classifier has been added to account for these correlations.

## 3.2 Personal tag profiles

A similar reasoning can be applied for tags given by one and the same user. It is not unreasonably to assume that users tend to reuse tags when posting questions on the same topics, and that the history of tags can thus improve future predictions. This information, combined with content-based analysis using topic modelling, could thus potentially improve the recall. For this reason, it is important that the user's tag history and tag correlation are modelled by *the same classifier that can capture the correlation between both*.

## 3.3 Social tag profiles

As an extra experiment, the effectiveness of \social" profiles was investigated as well. Based upon the tags that other users { the user's peers { gave in reply to a user's question, a profile was constructed containing the (normalized) frequency of tags given in the answers to a user's question. The reasoning behind this approach was the hypothesis that clusters of users tend to discuss the same topics. The results, however, were surprising: against expectations, the accuracy did not improve and in some cases even deteriorated with a few percentage points. A possible explanation for this behaviour is that tag usage is primarily personal, causing the proper tags { even if they are few { to be better predictors

| Characteristic | Value |
|---|---|
| # questions | 24748 |
| # distinct tags | 946 |

**Table 1: Characteristics of the "Ask Different" StackExchange dataset used for validating the performance of our methodology. More than 68% of all questions is annotated by at least one tag in the top-20. The 20th mostly used tag (mail.app) is used for only 3.1% of the questions.**

of future tags for a particular user. It seems, however, probable that this conclusion is not valid in other domains where the variety of tags is bigger than in a narrow folksonomy like StackExchange.

# 4. DISCUSSION

## 4.1 Dataset

A "real-life" data set from the \Ask Di erent" Apple-forum on StackExchange (periodically released for academic purposes), was used to validate the e ectiveness of our proposed methodology [1]. Table 1 summarizes its main characteristics. A 75%-25% split was used for obtaining a training and test set.

## 4.2 Single-label predictions

Previous work on predicting tags in Stack Exchange done by Stanley & Byrne achieved 65% accuracy using a model that took tag correlations into account [15]. Our two-stage model accomplishes similar results (65.7%) and lifts the accuracy to 73.8% for the top-20 tags when a user's history of tags is included.

## 4.3 Multi-label classification results

Table 2 summarizes the impact of the di erent topologies on the accuracy and exact match measures. The exact match (correct prediction all tags) improved from 18.2% to an impressive 54.3% for the top-20 and 73% for the 15th till 35th most frequently used tags. Though the accuracy does not seem to have been improved in case of "tag correlation", the recall improved from 42.7% to 52.4% on average for the top-20 tags, at the expense of a drop in precision of almost 26% from 93 to 67%. Tags that were previously undetected, are now identi ed (the classi er being overcon dent, predicting tags too often, causing a drop in precision). Though this seems a bad thing at rst sight, the increase in recall opens the pathway to better predictions when tag correlation is used in combination with a personal tag pro le. Including the history of past used tags during classi cation has the positive e ect of ltering out the incorrectly attributed tags (false positives) generated by the tag correlation. In other words, the negative e ect on precision is cancelled by e ect of the personal tag pro le: 81.9% instead of 67%. The recall climbs up as well to 72.6%. Fig. 2 summarizes these results.

## 4.4 "Rubbish" tags

When considering the distribution of tags, it appears that the mostly used tags do not follow a Zipf distribution. This gives support to our hypothesis that the very top tags are not used to communicate a topic, but rather to attract tra c. To test this hypothesis, the classi er was run on 3 datasets:

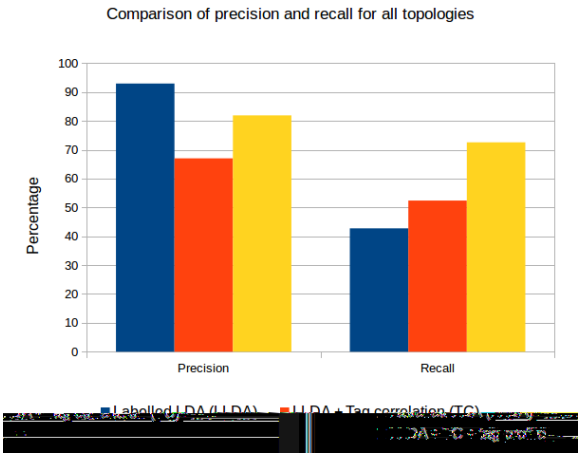


**Figure 2: Including the correlation between tags and a history of users' tags significantly improves the recall of tag predictions.**

(1) the top 20 of most used tags, (2) the 8th till 30th most used tags and (3) the top 15-35 as shown in Fig. 3. The exact match, by far the most strict measure for multi-label classi cation, increased from 54.3% to a impressive 78%. The fact that the prediction accuracy increases as tags become less common, suggests that the top keywords risk of being "over-used", making them lose semantic meaning.

## 4.5 Scalability

It is important to notice that the methodology does not scale well when the number of tags becomes very large, as the amount of single-label classi ers increases proportionally. This poses no real problem in a StackExchange setting as almost 70% of questions are tagged with the rst 20 tags only. In a di erent setting with virtually in nite tags, a hierarchical approach in which clusters of topics are identi ed through the presented methodology and individual tags are proposed by means of a collaborative ltering approach may prove a solution.

# 5. CONCLUSIONS

In this work, the impact of incorporating personal tag histories when attributing tags to users' posts was investigated. A hybrid method based on document classi cation (for content-based analysis) and a personal tag pro le (leveraging social features like in collaborative ltering) proved to improve the prediction accuracy signi cantly. Compared to the State-of-the-Art [15], our model accomplishes similar accuracy of 73.8% for the top-20 tags, compared to 65.7% for a di erent data set. More importantly, including a user pro- le, increased the *exact match* (all tags including the number of tags forecast correctly) considerably from 18.2% to 54.3%.

Interestingly enough, a social tag pro le (frequency of tags used by peers) had little or negative e ect on accuracy for the considered dataset. A potential explanation is that tag use is highly personal so that even a few tags from the user itself cancel the e ect of the social pro le. This nding should however be validated on other datasets. Also, it appeared that performance increased (from 54.3% to 78%) when the

| Topology | Acronym | Accuracy | Exact match |
|---|---|---|---|
| Labelled-LDA | L-LDA | 41.8% | 18.2% |
| L-LDA + tag correlation (TC) | LLDA-TC | 44.1% | 17.1% |
| LLDA-TC + personal tag pro le | LLDA-TC-PTP | 70.1% | 54.3% |

**Table 2: Accuracy and Exact Match scores for the different topologies, based on predictions of the top-20 tags.**

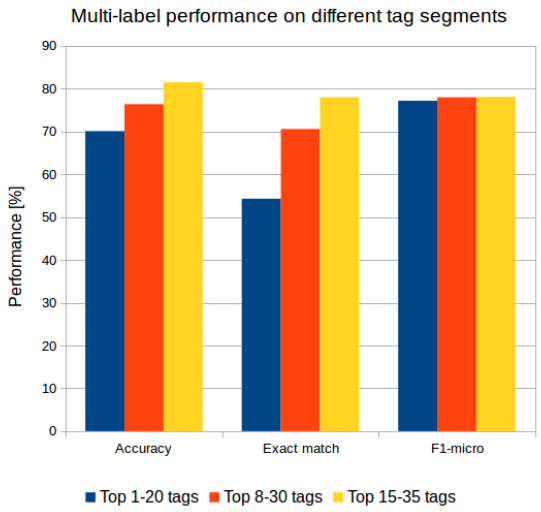


**Figure 3: Classification performance of the LLDA-TC-PTP topology improves as less common tags segments are considered. This observation supports the hypothesis that mostly used tags are over-used and "lose meaning".**

methodology was tested on less popular tags (15-35th), providing evidence for the "rubbish tag" hypothesis.

## Acknowledgments

The authors wish to thank Agentschap voor Innovatie door Wetenschap en Technologie (IWT) for their support.

## 6. REFERENCES

[1] "Ask Di erent" StackExchange Forum. http://apple.stackexchange.com/, 2015.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, J. A. Umass, B. A. Cmu, D. B. Cmu, A. B. Cmu, R. B. Cmu, et al. Topic detection and tracking pilot study nal report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[3] N. Bessis and F. Xhafa. *Next Generation Data Technologies for Collective Computational Intelligence*, volume 352. Springer Science & Business Media, 2011.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993{1022, 2003.

[5] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of documentation*, 35(4):285{295, 1979.

[6] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281{285. ACM, 1988.

[7] H.-N. Kim, A.-T. Ji, I. Ha, and G.-S. Jo. Collaborative ltering based on collaborative tagging for enhancing the quality of recommendation. *Electronic Commerce Research and Applications*, 9(1):73{83, 2010.

[8] M. Mezghani, C. A. Zayani, I. Amous, and F. Gargouri. A user pro le modelling using social annotations: a survey. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 969{976. ACM, 2012.

[9] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua. Learning to recommend descriptive tags for questions in social forums. *ACM Transactions on Information Systems (TOIS)*, 32(1):5, 2014.

[10] K. Nishida and K. Fujimura. Hierarchical auto-tagging: Organizing q&a knowledge for everyone. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1657{1660, New York, NY, USA, 2010. ACM.

[11] D. Quercia, H. Askham, and J. Crowcroft. Tweetlda: supervised topic classi cation and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 247{250. ACM, 2012.

[12] V. S. Rekha, N. Divya, and P. S. Bagavathi. A hybrid auto-tagging system for stackover ow forum questions. In *Proceedings of the 2014 International Conference on Interdisciplinary Advances in Applied Computing*, ICONIAAC '14, pages 56:1{56:5, New York, NY, USA, 2014. ACM.

[13] A. K. Saha, R. K. Saha, and K. A. Schneider. A discriminative model approach for suggesting tags automatically for stack over ow questions. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 73{76, Piscataway, NJ, USA, 2013. IEEE Press.

[14] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259{266. ACM, 2008.

[15] C. Stanley and M. D. Byrne. Predicting tags for stackover ow posts. In *Proceedings of ICCM*, volume 2013, 2013.

[16] H. Xie, Q. Li, X. Mao, X. Li, Y. Cai, and Y. Rao. Community-aware user pro le enrichment in folksonomy. *Neural Networks*, 58:111{121, 2014.