Group recommendations are presented to users along with explanations about the reasons that the particular items are being suggested. Explanations are given as text using a template mechanism.

A main problem of this approach is to identify the most similar users for each user in the group. A solution that involves no pre-computation requires computing the similarity measures between each user in the group and each user in the database. To avoid exhaustively searching for similar users, we perform some pre-processing steps offline. In particular, we propose building clusters of similar users, considering as similar those users that have similar preferences. To partition users into clusters we employ a bottom-up hierarchical agglomerative clustering algorithm. Initially, our algorithm places each user in a cluster of his own. Then, at each step, it merges the two most similar clusters. The similarity between two clusters is defined as the minimum similarity between any two users that belong to these clusters (max linkage). The algorithm terminates when the similarity of the closest pair of clusters violates a user similarity threshold δ. Ideally, the most similar users to a specific user are the members of the cluster that the user belongs to. Recommendations are computed based on the preferences of these cluster members. Figure 1 shows a high level representation of the architecture of our system.

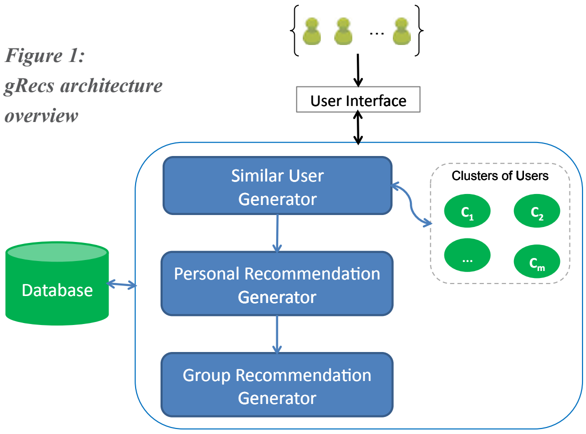

Figure 1: gRecs architecture overview

Our results show that employing user clustering considerably improves the execution time, while preserving a satisfactory quality of recommendations [1]. To deal with the high dimensionality and sparsity of ratings, we envision subspace clustering to find clusters of similar users and subsets of items for which these users have similar ratings.

We designed and developed the gRecs system at the Norwegian University of Science and Technology in Trondheim, Norway, funded by the ERCIM "Alain Bensoussan" Fellowship Programme, in collaboration with Irini Ntoutsi and Hans-Peter Kriegel from the Ludwig Maximilian University of Munich, Germany.

Reference:

[1] I. Ntoutsi, K. Stefanidis, K. Nørvåg and H-P. Kriegel: "Fast Group Recommendations by Applying User Clustering", in proc. of ER 2012.

Please contact:

Kostas Stefanidis, Kjetil Nørvåg, NTNU, Norway
E-mail: kstef@idi.ntnu.no, noervaag@idi.ntnu.no

Utility-Theoretic Ranking for Semi-Automated Text Classification

by Giacomo Berardi, Andrea Esuli and Fabrizio Sebastiani

Researchers from ISTI-CNR, Pisa, have addressed the problem of optimizing the work of human editors who proofcheck the results of an automatic text classifier with the goal of improving the accuracy of the automatically classified document set.

Suppose an organization needs to classify a set of texts under a given classification scheme, and suppose that this set is too large to be classified manually, so that resorting to some form of automated text classification (TC) is the only viable option. Suppose also that the organization has strict accuracy standards, so that the level of accuracy that can be obtained via state-of-the-art TC technology is not sufficient. In this case, the most plausible strategy to follow is to classify the texts by means of an automatic classifier (which we assume here to be generated via supervised learning), and then to have a human editor proofcheck the results of the automatic classification, correcting misclassifications where appropriate.

The human editor will obviously inspect only a subset of the automatically classified texts, since it would otherwise make no sense to have an initial automated classification phase. A software system could actively support the human editor by ranking, after the classification phase has ended and before the inspection begins, the automatically classified documents in a such a way that, if the human editor inspects the documents starting from the top of the ranking and working down the list, the expected increase in classification accuracy that derives from this inspection is maximized. We call this scenario "semi-automated text classification" (SATC).

A common-sense ranking method for SATC could consist in ranking the automatically classified texts in ascending order of the confidence scores generated by the classifier, so that the top-ranked documents are the ones that the classifier has classified with the lowest confidence [1]. The rationale is that an increase in accuracy can derive only by inspecting misclassified documents, and that a good ranking method is simply the one that top-ranks the documents with the highest probability of misclassification, which (in the absence of other information) we may take to be the texts which the classifier has classified with the lowest confidence.

We have recently shown [2] that this strategy is, in general, suboptimal. Simply stated, the reason is that, when we deal with imbalanced TC problems (as most TC problems indeed are [3]) and, as a consequence, choose an evaluation measure - such as F1 - that caters for this imbalance, the improvements in effectiveness that derive from correcting a false positive or a false negative may not be the same.

We have devised a ranking method for SATC that combines, via utility theory, (i) information on the probability that the document is misclassified, and (ii) information on the gain in overall accuracy that would derive by proofchecking it.

We have also proposed a new evaluation measure for SATC, called Expected Normalized Error Reduction (ENER). Since different users will inspect the ranked list down to a certain “inspection depth”, ENER uses a probability distribution over inspection depths as a parameter. ENER measures then the expected value (over this probability distribution) of the reduction in error that inspecting a ranked list down to the specified depth would bring about.

We have used ENER as the evaluation measure for our experiments, which we have run on a standard text classification dataset. The results show that, with respect to the common-sense baseline method mentioned above, our utility-theoretic ranking method is substantially more effective, with computed improvements ranging from +16% to +138%.

The approach we present is extremely general, since it applies straightforwardly to cases in which evaluation measures different from F1 are used; multivariate and non-linear evaluation measures can be handled too, provided they can be computed from a standard contingency table. By using our method, it is also easy to dynamically provide the human editor with an estimate of how accurate the classified set has become as a result of the proofchecking activity.

References:
[1] A. Esuli and F. Sebastiani: “Active Learning Strategies for Multi-Label Text Classification”, in proc. of ECIR 2009, Toulouse, FR, 2009, pp. 102-113

[2] G. Berardi, A. Esuli, and F. Sebastiani: “A Utility-Theoretic Ranking Method for Semi-Automated Text Classification”, in proc. of ACM SIGIR 2012, Portland, US, pp. 961-970

[3] H. He and E. Garcia: “Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering”, 21(9), 1263-1284.

Link:
http://nmis.isti.cnr.it/sebastiani/Publications/SIGIR12.pdf

Please contact:
Fabrizio Sebastiani, ISTI-CNR, Italy
Tel: +39 050 3152 892
E-mail: fabrizio.sebastiani@isti.cnr.it

A Radio Telescope of the Superlative

by Ton Engbersen

The worldwide community of Radio-Astronomy has envisioned building a very large, highly sensitive radio telescope partly in South Africa and partly in Australia by 2020. The total effective area of this radio telescope should approach one square kilometer and therefore it is called the Square Kilometre Array (SKA). The SKA instrument is expected to generate Exabytes of data per day which need to be processed and reduced, such that approximately 1 Petabyte per day is left to be stored for later use by Radio Astronomers.

Current expectations for the SKA are that the low frequency array (70 – 450 MHz) and the initial mid frequency (450 – 3000 MHz) will each comprise about 500,000 antenna elements while the high frequency array (3 – 10 GHz) will consist of approximately 3000 dishes. A quick calculation assuming no beamforming before Nyquist sampling results in 3.5 10^{15} samples/s or 300 ExaSamples per day (assuming 24 hour operation). Processing this is clearly beyond the capabilities of even the fastest supercomputers one can envision by 2020. The streaming and real-time nature of the SKA makes it unlikely that supercomputers are ideally suited for this application, like in LOFAR [1]. A significant research and development effort is therefore needed. For IBM, with our focus on future Big Data and Big Data analytics, this is a highly interesting field of research: it promises to make analytics low cost and energy efficient. We have named the project DOME after the protective astronomical telescope covering.

DOME

A five-year, 33 million Euro project has been defined between IBM Research – Zürich and ASTRON, funded by the Dutch Ministry of Economic Affairs, Agriculture and Innovation and the Province of Drenthe, The Netherlands. The objective is to investigate novel exascale computing technologies and concepts, with a focus on energy-efficient data processing, data storage, and nano-photonics at a fundamental level. In addition, the DOME project will collaborate with Small and Medium Enterprises and other academic partners in the Netherlands to stimulate economic activity through supporting the development and testing of new high-performance computing applications.

Research Projects

In DOME, seven research tracks are defined:

1. Algorithms and Machines: The goal is to design a whole-system bounds framework enabling system-design space exploration in the early phases of the SKA implementation and thus guide the design decisions for platforms which will hold future exascale systems. A methodology already in development in the IBM Laboratory in Zürich forms the basis: analytical models and equations tie application properties, device technology and compute architecture trends together to arrive at predictions of performance[2], power and hardware cost.