Data mining, interactive semantic structuring, and collaboration: A diversity-aware method for sense-making in search

Mathias Verbeke, Bettina Berendt and Siegfried Nijssen

Department of Computer Science, K.U. Leuven, B-3001 Heverlee, Belgium;
http://www.cs.kuleuven.be/~berendt

Abstract. We present the *Damilicious* method and tool which help users in sense-making of the results of their literature searches on the Web: on an individual level, by supporting the construction of semantics of the domain described by their search term, and on the collective level, by encouraging users to explore and selectively re-use other users' semantics. We use a combination of clustering, classification and interactivity to obtain and apply diverse semantics of thematic areas, and to identify diversity between users. This tool can help users take different perspectives on search results and thereby reflect more deeply about resources on the Web and their meaning. In addition, the method can help to develop quantitative measures of diversity; we propose diversity of resource groupings and diversity of users as two examples.

Keywords: Diversity aware classification and clustering technology; Diversity driven information aggregation technologies; Semantic clustering; Diversity aware search and semantic search;

1 Introduction

Search is one of the main applications on today's Web and other repositories, and it is supported by more and more advanced techniques. However, these techniques largely ignore an important component of search: the further processing of search-result sets that humans invariably undertake when dealing seriously with a list of documents provided by a search engine – and the diversity in which this is done by different people. An important form of such further processing is the grouping of result sets into subsets, which can be seen as investing an undifferentiated list of results with "semantics": a structuring into sets of documents each instantiating a concept (which is a subconcept of the overall query). On the Web, several search engines provide an automatic clustering of search results. However, regardless of how good the employed clustering algorithm is, a "globally optimal" clustering solution is generally impossible: There is no single best way of structuring a list of items; instead, the "optimal" grouping will depend on the context, tasks, previous knowledge, etc. of the searcher(s).

On a truly Social Web, users should be empowered to see and work with a diversity of possible structurings and their associated sense-makings. A prerequisite for such diversity-aware tools is that users are able to perform individual sense-making, structuring search result sets according to their needs and knowledge. The problem that we study is thus how to make users aware of alternative ways to group a set of documents. The approach that we take is to provide users with a tool to cluster documents resulting from a query. In this tool, *Damilicious* (DAta MIning in LIterature Search engines), the user can, starting from an automatically generated clustering, group a search result document set into meaningful groups, and she can learn about alternative groupings determined by other users. To **transfer** a clustering of one set of documents to another set of documents, the tool learns a **model** for this clustering, which can be applied to cluster alternative sets of documents. We refer to this model as the clustering's **intension**, as opposed to its **extension**, which is the original, unannotated grouping of the documents. This approach supports various measures of diversity. Such measures can be used to make recommendations and present new, possibly interesting viewpoints of structuring the result set. The domain of literature search was chosen to ensure a high level of user interest in reflection and meaningful results; the methods can easily be transferred to other domains.

The paper is structured as follows: Section 2 gives a brief overview of related work. In Section 3, we introduce the method of semantic clustering and classification. In Sections 4 and 5, we describe the individual and collaborative uses of the tool. Section 6 concludes with an outlook.

2 Background and related work

Our work extends theoretical and practical results from various areas.

To present search results in a more meaningful way, search engines such as `www.clusty.com`, `www.kartoo.com` or `search.carrot2.org` **cluster search results** based on distance measures on documents. Subgroups are labelled based on top keywords or phrases. However, users have no possibility of improving on the quality or context-relatedness of the results. (**Social-tagging applications** such as `www.delicio.us`, `www.citeulike.org` or `www.bibsonomy.org` are in a sense the opposite: Users are completely free in grouping and labelling resources they have found on the Web, but in general the only help they obtain from machine intelligence are equality matches with other users' tags or tag proposals.)

In knowledge engineering, these problems are addressed by techniques for **semi-automatic ontology learning** approaches where system-derived clusters can be modified and annotated by users, e.g. in the tool [9]; see also [7]. In CiteSeerCluster [3], we built on these ideas to improve the search for scientific literature. This tool employs textual clustering methods and builds on established bibliometric analyses and clustering algorithms, cf. [11]. Users create, modify and annotate groupings of the documents returned by a query, starting from an automatic clustering, and they engage in sense-making of the domain.

CiteSeerCluster has two limitations. First, each search episode stands alone: The results (document groups as extensions and automatically derived keywords as well as manually assigned labels and annotations as intensions of the concepts created by a structuring activity) can be saved and loaded, but not re-used or built on in new search episodes. Second, support for collaboration is therefore also limited to the exchange of combined episode intensions and extensions, i.e. "commented literature lists", but not intensions separately.

To overcome these limitations, we turned to recent work on **conceptual** and **predictive clustering** [4, 16]. The key idea that we use from this work is to a) form clusters of elements, then b) learn classifiers that reconstruct these clusters, c) validate these classifiers by investigating the reconstruction quality, and d) apply the classifier for further result sets.

Research on ontology re-use, in particular adaptive ontology re-use [14], investigates the modelling and mapping steps needed for re-using given ontologies, for whatever purpose. In contrast, we concentrate on re-use for grouping/classifying new objects, and on how to find the most suitable ontologies for re-use. Our aim is to derive, from this procedure, measures of **diversity** and to use these to support search. We build on the notion of diversity of result sets as used in recommender systems, e.g. [17], and information retrieval, e.g. [1], and on the notion of (cultural) diversity of people, cf. [10, 2].

3 Semantic clustering, classification, and interactivity

In this section, we describe the basic combination of clustering and classification methods for deriving and re-using a semantics of search that has been created by user and system together. We describe the system's general workflow and our measure of diversity between document groupings.

Clustering, classification and regrouping proceed through five steps. All steps, as well as the whole process, can be iterated:

1. Query: A user query is forwarded to a search engine or repository, and the list of results is shown in the Damilicious interface. We use the $CiteSeer^X$ repository[1] because of its broad coverage and rich structure, and also because it offers an OAI interface.[2] The output is a set of document IDs, document details and their texts.

2. Automatic clustering: The system clusters the documents with the Lingo algorithm[3]. For each of the resulting clusters, an intensional definition is calculated, where the set of intensional definitions of the clusters is the intensional definition of the clustering. These intensional definitions provide the criteria based on which the documents are clustered. They can thus be seen as a classifier, and can be used to classify new documents in the future.

3. Manual regrouping: Since the automatic clustering is only a suggestion for the structure, each user can move and delete documents (the extensions) between the different clusters.

4. Description of alternative ways of grouping: Different ways in which a result set can be grouped are compared and shown.

5. Transfer: A grouping solution can be used to group an enlarged or a different document set. This is done by applying the intension as a classifier. This re-use

[1] `http://citeseerx.ist.psu.edu/`

[2] an interface for harvesting metadata from separate repositories, see `www.openarchives.org`

[3] `http://project.carrot2.org`

can be done using groupings produced by the user herself ("individual", see Section 4), or produced by another user ("collaborative", see Section 5).

Measures of grouping similarity and diversity Step 4 of the workflow requires a measure of the similarity/diversity of groupings. We use normalized mutual information, because it has desirable properties: it has a value between 0 and 1, and it behaves well when calculated for clusterings with different numbers of clusters [13]. The normalized mutual information (specifically, NMI 4 [15]) of two groupings F, G is defined as $NMI(F, G) = (H(F) + H(G) - H(F, G))/\sqrt{H(F)H(G)}$, where $H(G)$ is the entropy of grouping G and $H(F, G)$ the joint entropy of F and G together.[4] For clustering algorithms like Lingo that can generate overlapping clusters, the probabilities used in the entropy calculations are normalized by division by the number of documents in each cluster. We treat NMI as a measure of the pairwise similarity of groupings, and $(1 - NMI)$ **as a measure of the (pairwise) diversity of groupings**.

One use of this measure was to choose the best clustering and classification algorithms for step 2 of the workflow. Our aim was a choice that reconstructs a "ground-truth" clustering as well as possible, i.e. has the highest NMI with it. We investigated two settings experimentally. The first is a 'sanity check': cluster document set D; learn a classifier, apply this classifier to D, compare the classes with the clustering. The second setting models the basic case of re-use, the issuing of the same query at a later time towards a grown database: cluster document sets D' and $D \subset D'$, to obtain clusterings G' and G. Learn a classifier from G, apply it to D', compare the classes with G'. We compared k-means and Lingo for clustering and top-10 TF.IDF words, Lingo phrases, and Ripper [6] rules for classification. The combination of Lingo with Lingo phrases gave the best results and is therefore implemented in the Damilicious user interface.

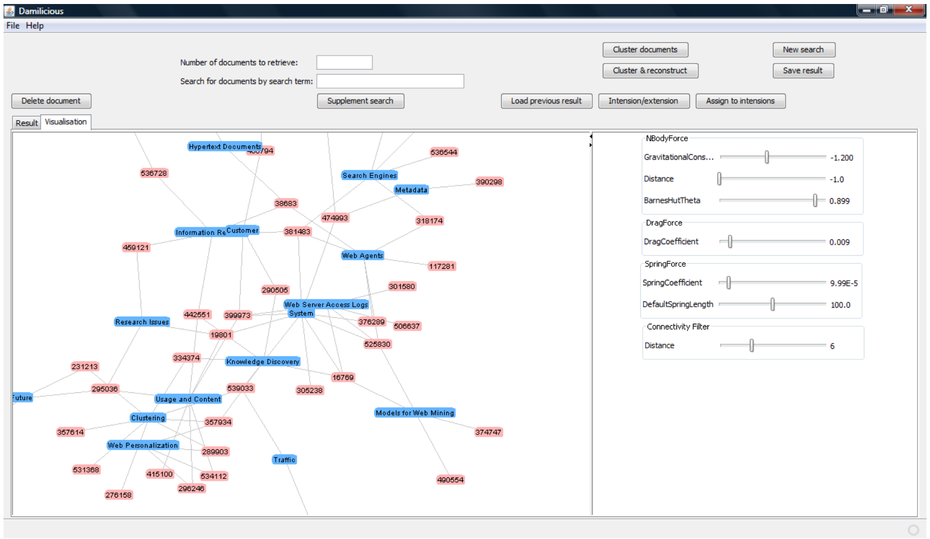
4 Individual sense-making: Interactive semantics

The use of Damilicious by one individual serves two goals: (1) to support individual sense-making and (2) to re-use one's own prior search episodes, more specifically the semantics created in them. (1) is realized through steps 1–3 of Section 3; results can be saved and re-loaded. Goal (2) is realized in step 5 by applying a classifier obtained in a previous search episode of the user.[5] Here, we illustrate step 4: the use of interactive cluster graphs to show the current (machine-generated or manually regrouped) extensions and intensions.

Example: Towards an individualised semantics of "web mining" Figure 1 illustrates how Damilicious supports conceptual thinking in search: As in other forms of semi-automatic ontology learning such as [9], the groups are viewed as concepts, with the extension of the concept given by the set of documents in the

[4] for clusters C_i and p the distribution of documents over them: $H(G) = -\sum_{C_i \in G} p(i) log_2 p(i)$ and $H(F, G) = -\sum_{C_i \in F} \sum_{C_j \in G} p(i, j) log_2\, p(i, j)$.

[5] The resulting 'stability' of document groups was highly appreciated by the student participants of a small user study.



Number of documents to retrieve:

Search for documents by search term:

NBodyForce

GravitationalCons... -1.200

Distance -1.0

BarnesHutTheta 0.899

DragForce

DragCoefficient 0.009

SpringForce

SpringCoefficient 9.99E-5

DefaultSpringLength 100.0

Connectivity Filter

Distance 6

Fig. 1. Damilicious: result screen

group. The intensions of the concepts are given here by the Lingo phrases; in [9, 3], they are given by a combination of top TF.IDF terms and manually assigned labels. The layout ensures that large, 'important' clusters that overlap with many others are shown in the centre, while small and 'isolated' clusters move to the margins. In the example, the system solution has identified "Web agents" as well as "Web personalization" as subfields of the field specified by the query, "Web mining". Unavoidably, in this automatic solution the top content area is not split up into subgroups by the same criterion; thus for example, "Clustering" (a method which could well be used for building agents or doing personalization) is another cluster. Users can improve on this structure: by moving and deleting documents between groups, they can modify the groupings to better reflect their intentions, background knowledge, interests, etc.

5 Collaborative sense-making: Diversity

The collaborative use of Damilicious serves two goals: (1) to get an overview of user diversity and to localise oneself in this space, and (2) to re-use other users' semantics. (1) builds on steps 1–3 of Section 3. Goal (2) is realized in step (5), more specifically, in the transfer of a foreign model to a new search episode. In this section, we illustrate step 4: the use of a measure and description of user diversity by a semantic space of users. First, the measure is explained; then one possible depiction is described; an alternative is to use a series of cluster graphs as in the previous section.

A measure of user diversity Above, we have proposed to use $1 - NMI(F, G)$ as a measure of the diversity between groupings F and G. In our setting, the instance set is a set of documents returned by a search engine operating on a given state of a database, in response to a query q. The grouping is either the result of a clustering/classification algorithm A (with parameters set by a user or optimised according to some strategy), or it is a manual regrouping done by a user A. To emphasize these two determinants, we replace F (or G) by $gr(A, q)$.

From this, we derive the measure $gdiv(A, B, q)$ of the diversity of users A, B (with respect to the way in which they structure the search result of q) as

$$gdiv(A, B, q) = 1 - NMI(gr(A, q), gr(B, q)). \tag{1}$$

Based on this, a **measure of the (pairwise) diversity of users** $gdiv(A, B)$ can be defined by an aggregation over queries they both worked on. A simple example of an aggregation operator is the average:

$$gdiv(A, B) = avg_{q:\exists gr(A,q) \wedge \exists gr(B,q)}(gdiv(gr(A, q), gr(B, q)). \tag{2}$$

Example (contd.): Diversity in conceptualising "Web mining" To demonstrate the use of the user-diversity measure, we simulated a small user study. Five users searched for "web mining" and restricted Damilicious to retrieving 50 documents. User U0 did not change the system clustering. User U1 regrouped documents to produce a better fit of the document groups to the cluster intensions, with a total of five document regroupings. User U2 attempted to move everything that did not fit well into the remainder group "Other topics", and in addition performed some regroupings to achieve a better fit; this resulted in ten regrouping operations. User U3 had the reverse goal: In a total of five regrouping operations, she distributed documents from "Other topics" into matching real groups. User U4 pursued a very different strategy, regrouping by author and institution; she also performed five regrouping operations.

For $q =$ *"web mining"* and each pair of users A, B, $NMI(gr(A, q), gr(B, q))$ was computed. Applying equation 1, we obtain a 5×5 matrix of differences $gdiv(A, B, q)$. Values ranged from 0.52 to 0.58. Assuming q to be the only commonly worked-on query and applying equation 2, these differences are equal to $gdiv(A, B)$ and together form the user-diversity matrix $\boldsymbol{GDIV}$.

This matrix can be interpreted as giving rise to a "semantic space of users", in which users who group in similar ways are close to one another, but far from users who group differently. Neighbourhood and distribution in this semantic space can be regarded as indicators of diversity: pairwise diversity (distances) between users, as well as overall distribution of population diversity. Multi-dimensional scaling (MDS) is a popular way for visualising such distance-based semantic spaces, e.g. [9]. Figure 2 (a) shows an MDS visualization[6] of $\boldsymbol{GDIV}$.

At the bottom right-of-centre, user U0 is shown as "System" (the system solution). The locations of the other users in this space can be interpreted as follows. U2 produced the most different solution, possibly through her most active re-grouping. U1's regrouping differed more from the System's than U3's, possibly because U1 took a high-level semantic view in his attempt to produce better-fitting groups, while U3 was less ambitious in his focus on re-distributing documents (only) from the "Other Topics" cluster. Interestingly, U4's attempt

[6] coordinates generated with Talisman, `http://talisman.sourceforge.jp/mds`

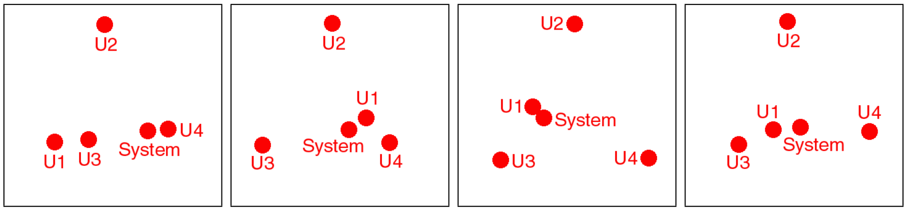


Fig. 2. Sample semantic spaces of users, showing their diversity: (a) web mining; (b) data mining; (c) RFID; (d) average of a,b,c.

to regroup by bibliographical data on author and institution (based on different data than the System that uses textual information from the document abstracts) resulted in the grouping most similar to the System's. It is possible that this is due to the fact that the same authors will use very similar wordings throughout their papers[7], such that the textual similarity is higher than that of other papers on the same subject (which was the focus of U1's, U2's and U3's restructurings). In addition to these pairwise observations, the results also show a simple form of population diversity: U1, U2 and U3, who had similar intentions when regrouping, appear in one "region" of this semantic space of user diversity (in the MDS solution of Fig. 2: to the left of System), clearly distinct from U4 who appears in a region of her own (to the right of System).

Figure 2 (b) and (c) show the diversities generated by the same user strategies on a semantically related query (b: data mining) and on an unrelated query (c: RFID). Interestingly, the more similar query (b) also appears to generate a more similar user diversity space. Averaging over all three queries, (d) shows the commonalities of all three: the similarity of U1 and U3 vs. U2 and U4.

6 Conclusions and outlook

In this paper, we have presented the Damilicious tool which helps users in sensemaking of the results of their literature searches on the Web: on an individual level, by supporting the construction of semantics of the domain described by their search term, and on the collective level, by encouraging users to explore the diversity of other users and to re-use other users' semantics.

Many open issues remain to be solved. They include (a) how to measure not only the diversity in clustering of search results, but also the diversity in results for different queries, and in general different result sets; (b) how to best present diversity of groupings and users (should all users be shown, only the nearest neighbour, or a ranking of neighbours? should other groupings rather than other users be shown? should "aggregated groupings" be shown, and how can these be defined? how should this functionality be integrated into a comprehensive environment supporting user and community contexts [12]? how can appropriate interpretations of MDS be supported and others discouraged [5]?); (c) how to motivate users to take advantage of regrouping and diversity functionalities

[7] cf. also the modelling, in CiteSeer(X), of specific forms of textual similarity in same-author papers: `http://citeseerx.ist.psu.edu/help/glossary`

(specific literature-search tasks may be beneficial, as suggested by the evaluation results of [3]); (d) how to find the best balance between similarity ("re-use solutions from users like me") and diversity ("re-use solutions from users unlike me so I can broaden my horizon"); (e) which measures of grouping diversity are most meaningful to users (our currently used measure NMI is purely extensional or instance-based; more intensional, structure-based or hybrid measures from the ontology matching field [8] could also be useful); and (f) which other sources of user diversity to leverage (for example, affiliation or research area). In future work, we aim to explore these issues theoretically and in user studies.

Acknowledgments S. Nijssen was supported by the Science Foundation – Flanders.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, & S. Ieong. Diversifying search results. In *Proc. WSDM '09*, 5–14, New York, NY, USA, 2009. ACM.
2. B. Berendt & A. Kralisch. From world-wide-web mining to worldwide webmining: Understanding people's diversity for effective knowledge discovery. In *From Web to Social Web*, LNAI 4737, 102–121. Springer, 2007.
3. B. Berendt, B. Krause, & S. Kolbe-Nusser. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing & Management*, in press.
4. H. Blockeel, L. De Raedt, & J. Ramon. Top-down induction of clustering trees. In *In Proc. of the 15th ICML*, 55–63. Morgan Kaufmann, 1998.
5. I. Borg & J. Lingoes. *Multidimensional Similarity Structure Analysis.* Springer, New York, 1987.
6. W.W. Cohen. Fast Effective Rule Induction. In *ICML*, 115–123. 1995.
7. D. R. Cutting, J. O. Pedersen, D. R. Karger, & J. W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. SIGIR*, 318–329. ACM, 1992.
8. J. Euzenat & P. Shvaiko. *Ontology Matching.* Springer, Berlin/Heidelberg, 2006.
9. B. Fortuna, D. Mladenic, & M. Grobelnik. Semi-automatic construction of topic ontologies. In *Semantics, Web and Mining*, LNCS 4289, 121–131. Springer, 2006.
10. G. Hofstede. *Cultures and organizations: software of the mind. Intercultural cooperation and its importance for survival.* HarperCollins, London, 1994.
11. F. Janssens, W. Glänzel, & B. De Moor. A hybrid mapping of information science. *Scientometrics*, 75(3):607–631, 2008.
12. C. Niederée, A. Stewart, C. Muscogiuri, M. Hemmje, & T. Risse. Understanding and Tailoring Your Scientific Information Environment: A Context-Oriented View on E-Science Support. In *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments*, 289–298, 2005.
13. D. Pfitzner, R. Leibbrandt, & D. Powers. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, 2009.
14. R. Stecher, C. Niederée, W. Nejdl, & P. Bouquet. Adaptive ontology re-use: finding and re-using sub-ontologies. *IJWIS*, 4(2): 198-214, 2008.
15. A. Strehl & J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Machine Learning*, 3:583–617, 2003.
16. B. Zenko, S. Dzeroski, & J. Struyf. Learning predictive clustering rules. In *In Proc. 4th Int. Worksh. on Knowledge Discovery in Inductive Databases*, LNCS 3933, 234–250. Springer, 2005.
17. M. Zhang & N. Hurley. Statistical modeling of diversity in top-n recommender systems. In *Proc. of ACM/WIC/IEEE WI-IAT'09*, 490–497. 2009.