Scalable Computing: Practice and Experience
Volume 9, Number 4, pp. 293–301. http://www.scpe.org
ISSN 1895-1767
© 2008 SCPE

DERIVING A LIGHTWEIGHT CORPORATE ONTOLOGY FORM A FOLKSONOMY: A METHODOLOGY AND ITS POSSIBLE APPLICATIONS

CÉLINE VAN DAMME*, TANGUY COENEN† AND EDDY VANDIJCK*

Abstract. Companies use company-specific terminology that may differ from the terminology used in existing corporate ontologies (e.g. Tove) and therefore need their own ontology. However, the current ontology engineering techniques are time-consuming and there exists a conceptual mismatch among developers and users. In contrast, folksonomies or the flat bottom-up taxonomies constituted by web users' tags are rapidly created. In this paper, (1) we present an approach that cost-efficiently derives a lightweight corporate ontology from a corporate folksonomy, (2) by means of a folksonomy dataset from a European company, we provide preliminary evidence that our suggested approach reflects the company-specific terminology, (3) we detect a number of possible applications for the company when implementing the presented methodology on a corporate folksonomy and (4) as an additional evaluation, we asked the company to briefly evaluate the results and possible applications.

Key words: ontology, folksonomy, company, applications

1. Introduction. It has been stated, e.g. in [24, 6] that ontologies improve the communication among humans or machines since they provide a shared understanding of a domain. This makes that ontologies are very useful for companies. For instance they can help to improve the communication between employees.

At this moment, there exist several corporate ontologies, for instance Tove [7] and Enterprise ontology [26]. These ontologies describe general concepts and relations related to enterprise and process modeling. We believe these kinds of ontologies may not be useful for every enterprise since companies have a corporate-specific terminology and consequently have their own concepts. In our opinion, an enterprise may need its own corporate ontology.

Building ontologies with the current ontology engineering techniques have disadvantages. First of all, it is a very time-consuming process [2] and secondly the actual users are not involved in the developing process. As a consequence there exists a conceptual mismatch between the developers and the actual users' vocabulary [11].

These disadvantages are not present in the relatively new categorization method called tagging and its resulting folksonomy. Following the Web2.0 paradigm, a growing number of websites incorporate a tagging/folksonomy mechanism. They allow users to refer to resources (bookmarks, pictures or scholarly publications) on the web with freely selected keywords or tags. The users are not restricted to a controlled vocabulary produced by a group of experts. Users can enter any words that enter their mind. This makes them active participators in creating new tags. Aggregating this user created meta data leads to a flat, bottom-up taxonomy, also known as a folksonomy.

Despite the strengths, tagging has its weaknesses: no conceptual meaning or hierarchical relations are added to the tags. As a consequence, tags have no synonyms or homonyms. Furthermore, specialized as well as general tags can be used to annotate the same resource [9, 10]. These weaknesses can be solved by (1)giving the users tools that enable them to add more information to their tags (e.g. cluster tags as on Delicious) [10] and/or (2) trying to generate more information on the tags by employing text mining, statistical techniques and asking additional feedback from the community [4].

The last few years, we observe a growing attention of the semantic web community for tagging and its resulting folksonomies. At the one hand, we observe researchers that try to enrich the flat ambiguous tags with existing online resources (e.g. Google, Wordnet, existing ontologies) [22] and on the other hand, there are researchers that consider this user created meta data as a valuable source to develop ontologies [4].

In this paper, we argue that cost-efficiently deriving a lightweight ontology from a folksonomy is also applicable to a corporate folksonomy. We regard a lightweight ontology as the simplest form of an ontology: an ontology where only one relation is included or a taxonomy as described by [25]. We propose a 6-step approach which includes several techniques such as the Levenshtein metric, co-occurrence, conditional probability, transitive reduction and visualization. Although, some suggestions have already been made on how a corporate ontology can be built from a corporate folksonomy [3], no research results have been published so far. We implemented our approach on a corporate folksonomy of a large European distribution company in which Dutch and French are the two official company languages. We obtained the simplest form of an ontology, a lightweight

*MOSI, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium, {cvdamme, eddy.vandijck}@vub.ac.be
†STARLab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium, Tanguy.Coenen@vub.ac.be

ontology, and visualized it with the open source tool Graphviz (`http://www.graphviz.org/`). By means of the generated lightweight ontology, we were able to detect other possible applications than the one to improve the communication among the employees in the company. As an additional evaluation, we asked the company to evaluate the results and its applications.

The paper is structured as follows: we provide an overview of related work in section 2. In section 3, we discuss all the techniques of the methodology and explain how they can be integrated in our 6-step approach. In section 4, we elaborate on the corporate folksonomy dataset discuss the general results of applying our approach to the dataset. We describe possible applications of the approach for the company in section 5. Section 6 discusses our findings and presents future research. A conclusion is provided in section 7.

2. Related Work. At the time of writing, few papers have been written on discussing the use of folksonomies in a company. The authors in [17] present a social bookmarking tool, called Dogear, that lets employees tag their bookmarks from the corporate intranet and the World Wide Web. The advantages of collaborative tagging in the enterprise are discussed in [12]. The authors suggest that tagging can be used as an expert location tool that facilitates the process of organizing meetings with experts in the company. Tags are a reflection of people's interest and/or knowledge and can as a consequence be seen as a tool to detect experts and their domain of expertise.

However, the authors in [17, 12] do not explain how to make the tags less ambiguous nor turning them into an ontology. This is discussed in [3]. The authors propose to derive a CRM or Customer Relationship Management ontology from a corporate folksonomy. They suggest an integrated visual approach that integrates text mining techniques, tags and user feedback. Each time the employee adds a message or note to the CRM system, tags are required. At the same time, automatic keywords are detected based on the tf-idf score. The tf-idf score is calculated by multiplying the word's document frequency by the logarithm of its inverse document frequency in the set of relevant company documents. The higher the score, the more descriptive the keywords are [20]. In a first phase the user has to indicate whether there exists a relationship between the tags and the keywords with the highest tf-idf score. The relationship has to be specified in a second phase. In this approach, the human effort as well as the implementation time is very high. We also have to point out that the proposed approach still has not been tested.

Literature on folksonomies enrichment or turning folksonomies into ontologies is currently more common in the domain of the World Wide Web. In [21] tags of the photo-sharing site Flickr (`http://www.Flickr.com/`) were used in an experiment to induce a taxonomy, the simplest form of an ontology [25]. The approach of [21] is based on statistical natural language processing techniques where a subsumption or hierarchical relation was deducted. The authors of [22, 4] both suggest to include different techniques as well as the wealth of existing online web resources such as Wordnet, Wikipedia, Google, online dictionaries and existing ontologies. The authors in [22] present an approach to enrich tags with semantics to make it possible to integrate folksonomies and the semantic web. The authors use online lexical resources (e.g. Wordnet, Wikipedia, Google) and ontologies to map tags into concepts, properties or instances and determine the relations between mapped tags. However, the resources are tapped in one way (e.g. Wikipedia is used as spelling checker for tags) and the community is not involved to confirm the semantics obtained from existing ontologies and resources. Consequently, tags that reflect new concepts, relations or instances or new relations between tags are neglected. On the contrary, the opposite is suggested in [4]: ontologies are derived from folksonomies. Online lexical resources are suggested to be exploited in several ways. For instance Wikipedia is suggested as a spelling checker as well as a tool for finding concepts and homonyms. Furthermore, the authors suggest involving the community.

However, a corporate folksonomy differs from a folksonomy created on the World Wide Web. The users, their underlying motivations and the environment can be different. In case of a corporate folksonomy the user or employee is known and will not always tag voluntarily. An employee may be enforced to tag or may be given an incentive by the company. As a consequence, the amount of additional feedback asked from the users to create a lightweight ontology should be reduced. Labor costs are very high and therefore the number of employees involved with the feedback process should be minimized. In contrast to web communities it is far easier to ask the cooperation of the community: community members have a different mindset than employees and are more willing to participate in additional processes. However, in most cases they are anonymous. Company-specific terminology is mostly used in a closed company environment which makes it hard to include web resources in the ontology construction process. The terminology may contain terms which have a specific meaning for only a small group of employees.

3. Methodology. In this section, we first describe the different techniques we implement in the 6-step methodology, motivate why we do not include other techniques or online resources yet, and then elaborate on how we integrate the selected techniques as a whole.

3.1. Overview of techniques.

3.1.1. Levenshtein metric. The Levenshtein metric is a text similarity metric which calculates the distance between two words. More specifically, it counts how many letters have to be replaced, deleted or inserted to transform one word into the other [13]. It is a valuable technique to verify the similarities of two tags. In order to calculate the distance, first all possible tag pairs have to be made. In [22] a threshold value of 0.83 is used to indicate that two tags are similar. Yet tests showed us that a threshold value of 0.83 excluded a number of similar tags. For instance, the Dutch nouns *fiets* and *fietsen* or *bicycle* and *bicycles* in English, express the same thing but do not agree in number. Both tags are the same and their Levenshtein similarity is lower than 0.83. We believe this technique should be employed at a lower threshold value, we suggest 0.65, and include human feedback. A representative employee that is very well aware of all the terminology used in the company can be asked to confirm or reject the similarity.

As a tag cleaning method, we prefer this one to the one often suggested in literature, stemming. A stemming algorithm reduces tags to their stems or roots. The algorithm removes suffixes and hereby e.g. reduces the words *linked* and *links* to *link* [19]. The algorithm includes rules that are language dependent. Company-specific language can be lost because of the stemming algorithm. These words can differ from the general spelling rules or they can be abbreviations. Some languages, such as Dutch, incorporate English words in the vocabulary without adjustments to the Dutch language.

When stemming algorithms are used, there should be a way to determine the language of the tags and whether it involves corporate-specific language.

3.1.2. Co-occurrence. Luhn [14] stated that the frequency of words in a text can be used as a technique to detect relevant keywords for a document. Later, researchers in the domain of computational linguistics have started to use the statistical technique co-occurrence, the occurrence of two words used together in a text, to cluster terms [18]. [15] used a methodology based on co-occurrence to select the keywords for a document without a corpus or set of related documents. The co-occurrence technique is also proposed in the literature on folksonomies [21, 22]. For each tagged resource all the tag pairs are determined. The tie strength between a tag pair is increased each time two tags are used together.

It is interesting to know which tags are often used together to have already an idea which terms are often used together.

3.1.3. Conditional Probability. A rule based on the conditional probability definition was proposed in [16, 21]. More specifically, the rule tries to find out whether one of the tags in the pair can be defined as broader and the other one as narrower term. By applying the definition of the conditional frequency, the conditional probability is calculated by dividing the co-occurrence of the tag pair by the frequency of the individual tags. Results vary between 0 and 1. The higher the result, the more the term is used in combination with the other term and consequently the more depended it is of the other term. When the difference between the two results exceeds a certain threshold value, in [21] the threshold value is set to 0.8, a subsumption relationship is found.

Finding an appropriate threshold value should be determined based on trial and error testing.

3.1.4. Transitive Reduction. In [21] the authors remove the roots that are logically above the parent nodes. However, we believe transitive reduction, a technique from graph theory, is far more interesting. Transitive reduction reduces the edges of a graph G to a graph G' by keeping all the paths that exist between the nodes in Graph G [1]. The edges are consequently removed because of the implied transitivity.

3.1.5. Visualization Techniques. The use of visualization is proposed in [3] to lower the barriers to participate in naming the relations between concepts. In literature, several approaches for visualizing tags and lightweight ontologies are described. In [27] CropCircles are suggested to help people understand the complexity of a class hierarchy. We hypothesize that visualizing the lightweight corporate ontology may facilitate the validation process of the approach.

3.2. Other Techniques and online resources. Of course, a lot of other techniques (e.g. clustering techniques) or online resources could be interesting to extend the ontology with more relationships.

In [22, 4] the use of online resources such as Google, Wikipedia, online dictionaries is suggested as additional mean. The resources are regarded as spelling checkers and as a mean for retrieving concepts. The company-specific terminology makes it hard to use some of the sources on the internet. For instance, a company had a *gara* tag, used as the abbreviation of the Dutch word *garage*. When using *gara* as a search term for Google, we did not find any link referring to the correct meaning of the term. On Wikipedia, we found a page describing the term, but the concept or description attributed to it was incorrect. On Wikipedia, *gara* is a Basque word and the name of a Spanish newspaper. This causes problems. We have to know whether the tag belongs to the specific terminology of the company or not. In order to find this out, human feedback is necessary. However, asking employees to verify the word's background can quickly become too time-consuming. Therefore, we decided not to include any web resources yet.

3.3. 6-Steps Approach . Based on the techniques discussed in previous section, we explain how they can be integrated into our 6-step approach to derive a corporate ontology form a corporate folksonomy.

3.3.1. Step 1: Selection of the Tags. First, we remove all the Dutch stop words (Based on the list available at `http://snowball.tartarus.org/algorithms/dutch/stop.txt`) and filter the messages with fewer than 2 tags. We then withdraw the less frequently used tags by ranking the tags in an absolute frequency. Although in the domain of automatic indexing upper as well as lower bounds are used to exclude non-significant words, we assume that removing the upper bound tags will remove important company-specific elements for our lightweight ontology [8].

3.3.2. Step 2: Clean the Tags. Since folksonomies do not restrict its user to use a controlled vocabulary or predefined keywords, tags are polluted (e.g. plural and singular tags) and need to be cleaned up. We use the Levenhstein similarity metric combined with human feedback.

Based on a trial and error method, we decide to take 0.65 as a threshold value. All the tag pairs that reach a Levenhstein similarity of 0.65 will be presented and when two keywords are similar, the user has to check the corresponding check button, as visualized in figure one.

Then, the tag with the lowest frequency will be replaced with the one with the highest frequency. We opt for this rule since we believe that the tag with the highest frequency determines how the word should be written by the wisdom of the crowds in the company [23].

In figure 3.1, there are 4 tag pairs checked as similar. The tags with the highest frequency are always on the left. In the case of the tag pair (*winkel winkels*) or (*shop shops*) translated into English, the tag *winkels* will be replaced with *winkel* in the database. Whereas the tag pair (*artikel1234 artikel1235*) will not be adjusted. Latter tag pair contains dissimilar tags because they express different article numbers.

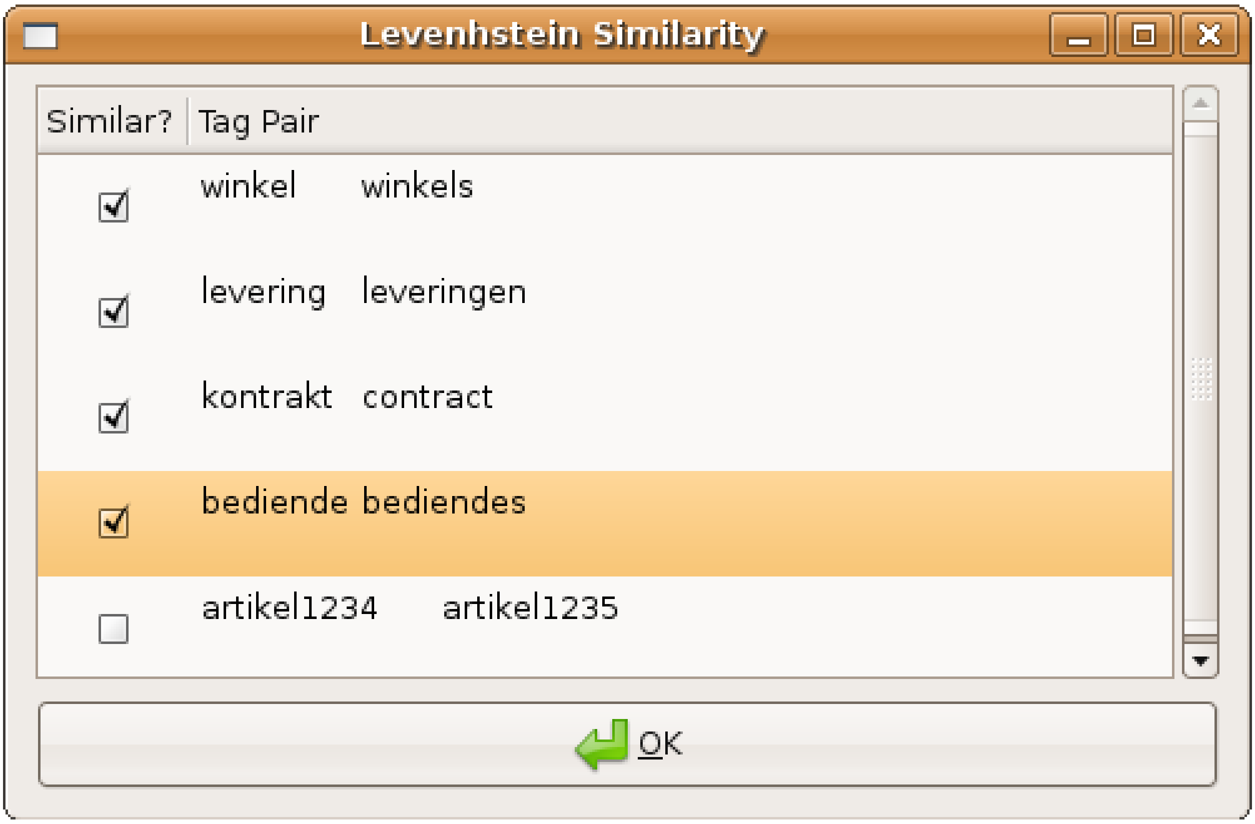
After the adjustment, we reselect the tags following the same procedure as described in the first step.

3.3.3. Step 3: Co-occurrence. For each message we make all the tag pairs. Then, we count the frequency of each unique tag pair. The more two tags are used together, the higher this frequency or co-occurrence value. Again, we decide to include only the ones with the highest frequency to find the most frequent relations.

3.3.4. Step 4: Finding Broader/Narrower Relations. We want to derive the simplest form of an ontology and therefore need to find the broader/narrower relations between the terms, for instance the relation between *animal* and *dog*. We apply the conditional probability function as described in previous section. Therefore, we divide the co-occurrence of the tag pair by the frequency of the tag itself. We did some manual tests deciding on 0.70 as the most appropriate threshold value. The higher the threshold value, the broader and the less deep the resulting ontology will be. For instance, when the tag pair *animal dog* occurs a 100 times and the frequency of both tags is respectively 500 and 120, we obtain the following results: animal = 0.2 and dog = 0.83. The tag dog exceeds the threshold value of 0.70 and therefore the relation between *animal-dog* can be considered as a broader narrower relationship.

3.3.5. Step 5 & 6: Transitive Reduction and Visualization. First, we apply the transitive reduction and then we visualize the remaining relations through Graphviz.

4. Dataset. In this section, we present the corporate folksonomy dataset and explain the results of applying our approach to this dataset.



Similar?	Tag Pair
☑	winkel winkels
☑	levering leveringen
☑	kontrakt contract
☑	bediende bediendes
□	artikel1234 artikel1235

Fig. 3.1. *Asking human feedback based on the Levenshtein metric*

4.1. Description corporate folksonomy. We have implemented our approach in a large European distribution company with headquarters in Belgium in which Dutch and French are the two official company languages. The company employs more than 15.000 people across Europe.

Tagging has been used on all their communication messages for more than 20 years. Messages such as letters and faxes that are not sent electronically are manually scanned, tagged and archived into an information system. Tags replace the subject line of the message. Tagging is completely integrated in the corporate culture. The messages can be created manually, automatically and semi-automatically. The automatic and semi-automatic messages have default tags. In case of semi-automatic messages, the author has to add complementary tags. Manually created messages require user created tags.

Initially, tags were introduced to solve the information retrieval problem since full text search engines were not available at the time. Tagging has remained part of the communication messaging system. However, the ambiguity of the flat tags and the information overload obstructs the search process. The company introduced some tag rules such as a minimum number of tags, no stop words, no plurals and no conjugated verbs, but only a minority of the employees in the company obeys all these rules.

Even though the tagging system at this company is somewhat different from current web-based tagging practices, the 20-years worth of tagged messages represented a real opportunity to test out the approach in a real-life case. Such cases are rare, as not many organizations have adopted tagging in a way which allows the analysis of a large body of tags. Tagging is so widely adopted and part of the corporate culture we believe the tags can be made to represent a non-toy lightweight ontology.

4.2. Tag datasets. In 2006, more than 8.000.000 messages were created and roughly 60.000.000 tags in total were used. 91% of the messages are created by Dutch speaking employees.

Due to the large size of the dataset and limited computer power, we decided to make a selection of the tags. We focused our analysis on the tags added to Dutch messages. More specifically, we analyzed 2 different message types individually: quick internal messages and notes since these are often used message types in the company.

As we discuss in the following paragraphs, we split the dataset into two sets and applied the 6-steps approach to tags annotated to quick internal and notes message types from both datasets.

4.2.1. Tag dataset 1: tags from automatic, semi-automatic and manual messages. At the beginning, we were not able to make a distinction between tags from automatic, semi-automatic and manual messages since a unique field to filter out the manual ones is not stored by the company. Therefore, the first tag dataset consisted of tags from the automatic, semi-automatic and manual messages.

Some information systems in the company can send automatic messages to the employees to inform them on certain issues, for instance an employee confirms to be present at a certain meeting and the system automatically sends a message to the person who organized the meeting. Tags are automatically generated and added to the message. In the case of semi-automatic messages, a message is based on an existing template including a list of tags that have to be extended. Whereas in the case of manual messages, the message as well as the tags are manually created.

We applied the approach to this dataset and after tag cleansing, we selected a group of tags (approximately 150) with a very high frequency (between 5000 and 147.000) to grasp the meaning and interrelations of these frequently used tags. We did the same for the selection of tag pairs.

In figure 4.1, a part of the obtained lightweight ontology of the quick internal messages is visualized. We renamed the top level node “name_of_shop” to guarantee the anonymity of the company.

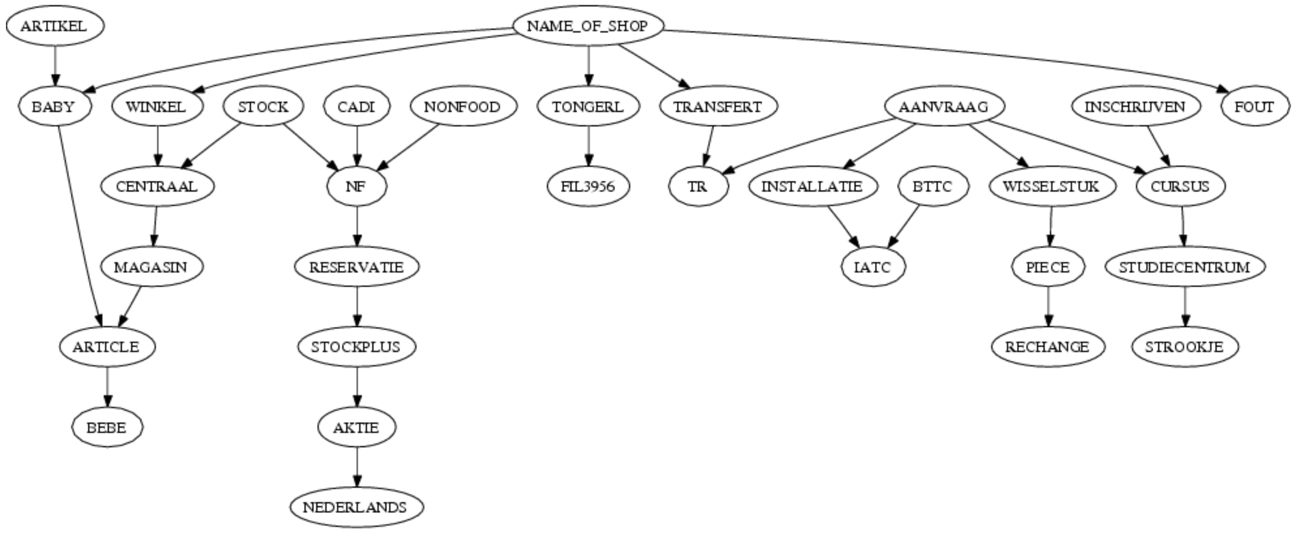


FIG. 4.1. *Partial results obtained from analyzing the quick internal messages from dataset 1*

4.2.2. Tag dataset 2: tags from manual messages. After presenting and discussing former results at the company, we realized it would be interesting to filter out the manual created tags. Apparently, many messages are automatically created and therefore partially influence the results received through previous dataset.

Based on the additional information given by the company, we were able to write a small script that allows us to make a distinction between the different kinds of messages. In total there are around 7.340.000 Dutch messages created in 2006. 72% of them are automatically created, 23% manually and 5% semi-automatically.

The same steps of the approach were applied to this dataset. Again, we selected a set of tags which have a frequency of more than 1.000, and employed the same threshold values as described in the approach. Finally, we received the result displayed in figure 4.2.

4.3. Discussion of Results. When visually comparing the output of the two message types, we notice that the 2 generated lightweight ontologies contain different terms. This means that the tag usage between the two message types differs. Consequently, we will need to find a way to map the different partial results into a complete ontology.

We notice that we have captured other relations than merely broader/narrower or *a kind of* relations. For instance the relation between the tags *name of shop* and *baby*, can not really be considered as *a kind of* relation

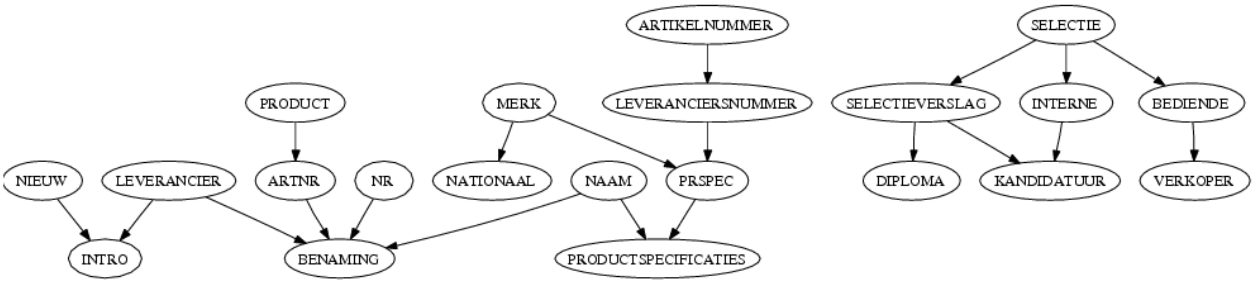


Fig. 4.2. *Partial results obtained from the message type "Notes" from the second dataset*

but more like a *is related to* relation. It provides more information regarding a stock item of the shop. Therefore, it would be interesting to find a way to capture these different kinds of relations and also check whether we may still apply transitive reduction.

We also observed that the graphs, as in figure 3.1, include some tags corresponding to the French language such as *article*, *bebe*, *magasin*, *piece*, *rechange*. When having a closer look at the data set, we noticed that there are some bilingual messages with bilingual tags. The tags can not be directly filtered from the database since there is no unique identifier. Looking at the results, we observed a pattern: the same tag relation exists between the Dutch and French tag pair e.g. in figure 3.1 (artikel, baby) and (article, bebe). We also observed this in the other results which are not visually included in this paper.

Tests with the Levenhstein metric, revealed that we can eliminate some French tags due to the close similarity among both languages e.g. *factuur* in Dutch and *facture* in French. In this way, the Levenshtein metric can reduce the pollution by French tags.

By applying our approach to these tags, we have reduced their tag's weaknesses as described in the first section. We now know with which other terms tags are mostly used together, for instance the tag *fout* is often used together with the tag *name_ of_ shop*. Pollution such as singular and plural tags is filtered out.

Since some parts of the obtained lightweight ontology are logically interpretable, we briefly verified the results by presenting them to the IT-director and the communication system's analyst of the company. They verified the results by looking at the visualizations and checking the tags in the communication system messaging system. They both confirmed that it reflects the company's terminology. Therefore, we concluded that the approach would be valuable to improve the communication among the employees. It visualizes how terms are often used together. When applying the approach on the tag dataset of every department, we should be able to compare the terminology of the different departments.

5. Possible Applications. Ontologies can be used to improve the communication in the company as motivated by [24, 6]. However, we believe that the methodology which we presented in this paper can be used for other applications than merely improving the communication among the employees in the company. The fact that the methodology is based (1) on the analysis of meta data or tags generated by the employees in the company and (2) the tagging process of the company under study is completely integrated with the actual business processes, generates a broad overview on the activities taken place over a certain time period.

As we will explain in the next paragraphs, we believe the visualization obtained from the approach could be used as a decision tool for management, follow-up tool for new terminology and as a tool for the creation of new teams.

5.1. Decision Management Tool. We believe that our methodology of building a visual lightweight corporate ontology from a folksonomy can be considered as a kind of business intelligence tool. Business intelligence aims at discovering interesting information based on analyzing the existing data in the company in order to improve the decision making process and generate a competitive advantage [5].

By observing figure 1, we noticed two remarkable relations. On the one hand, we saw that there exists a link between the *name of shop* (we renamed this tag to guarantee the anonymity of the company) and the tag *fout* or *mistake* in English. On the other hand, we found a relationship between the *name of shop* and the tags *Tongerl* and *Fil3965*. The tag *Tongerl* is used as the abbreviation for a Belgian city and *Fil3965* is the ID of one of the shops. The first mentioned relationship could be a signal that something is wrong and that the relationship between these tags should be further investigated. The latter one could indicate that the shop *Fil3965* has high sales revenue or high customer's complaints. By taking the time factor into account, these results could be compared over different time periods. Therefore, the approach presented in this paper might be an interesting tool for high-level managers in the company. High-level managers are more focused on higher level company's issues such as corporate strategy and are not always aware of all the things that are going on in the company. The visualization of the lightweight ontology obtained through our approach could support them in their daily work and help them in decision making. Therefore, we regard it as a kind of tool for decision making or a sort of add-on for an existing business intelligence tool.

5.2. Follow-up Tool for new Terminology. The proposed approach could be valuable as a follow-up tool for new corporate terminology. It reveals how new terms are utilized and interpreted. In the case of company acquisition, such an approach could be very interesting. When a company gets acquired by another company, the acquired company will have to apply new terminology to improve the communication process between both of them. Again, the time factor can be included in the process to evaluate and compare the results.

5.3. Creating Teams. When new teams have to be set up, the approach might be helpful to choose the most appropriate employees. This visualization shows how tags are combined with other ones. By selecting all the terms that are related to a certain word, the corresponding employees could be selected for the creation of a new team. Of course, social networking techniques [16] which can be used to cluster employees based on shared tags, can be used as an additional technique to find employees.

6. Discussion and Future Research. Next to briefly validating the approach by presenting the results to the IT-director and communication system's analyst of the company, we also discussed the possible applications of the approach. In their opinion, the first and third application benefit would be most interesting to their company. They even suggested a visual search tool as an additional application. Such as tool could be an extension of the suggested management tool. When the manager finds an interesting hierarchical relation or cluster, he should be able to click on it to retrieve the corresponding messages.

We plan to expand our tests to other message types to verify the applications which we deduced from our current results. In addition, we should set up focus groups with employees of the company where the results and the possible applications can be extensively discussed. The approach should be further extended and include more techniques and algorithms such as clustering techniques. In this way, more relations might be included in the ontology.

A threshold value that determines the minimal optimal frequency of a certain tag to be taken into account when applying our methodology should also be found.

When taking tags into account for business intelligence applications, the quality of the tags, becomes an important issue. Tagging does not restrict its users to use a predefined controlled vocabulary, they are free to use whatever tags or keywords they like. Since no control mechanism is included, there is no certitude regarding the quality of the tags. Therefore, metrics to automatically detect high quality tags becomes a real necessity.

Further, we will try to find a method to map the ontologies obtained by applying the approach to different message types. However, we believe a cost-benefit analysis should also be built-in in the approach to evaluate whether a more extended version of the ontology will generate the necessarily return on investment. Currently, the approach minimizes the human input and in this way a lightweight-ontology is cost-efficiently derived from the corporate folksonomy.

7. Conclusion. Companies need a corporate ontology because it can improve the communication among the employees. Since current ontology engineering techniques have some disadvantages, we proposed a new ontology engineering technique based on corporate folksonomies. It is a 6-step approach to turn a corporate folksonomy into a lightweight corporate ontology. By means of a corporate folksonomy, we applied our approach to an existing corporate folksonomy dataset. Based on a first small validation we concluded that the obtained lightweight ontology reflects the company's terminology and might help to improve the communication among the employees. We also deduced a number of possible applications for a company: decision tool for management, follow-up tool for new terminology and as a tool for the creation of new teams.

REFERENCES

[1] A. V. Aho, M. R. Garey, and J. D. Ullman, *The transitive reduction of a directed graph*, SIAM J. Comput., 1 (1972), pp. 131–137.

[2] E. P. Bontas and C. Tempich, *Ontology engineering: A reality check*, in The 5th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE2006), R. Meersman, Z. Tari, et al., eds., vol. 4275 of LNCS, Montpellier, France, Nov 2006, Springer, pp. 836–854.

[3] C. V. Damme, S. Christiaens, and E. Vandijck, *Building an employee-driven crm ontology*, in Proceedings of the IADIS Multi Conference on Computer Science and Information Systems (MCCSIS): E-society2007, 2007.

[4] C. V. Damme, M. Hepp, and K. Siorpaes, *Folksontology: An integrated approach for turning folksonomies into ontologies*, in Proceedings of Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007), 2007, pp. 57–70.

[5] P. Davies, *Intelligence, Information Technology, and Information Warfare.*, Annual Review of Information Science and Technology (ARIST), 36 (2002), pp. 313–52.

[6] D. Fensel, *Ontologies: : A Silver Bullet for Knowledge Management and Electronic Commerce*, Springer, 2003.

[7] M. S. Fox, *The tove project towards a common-sense model of the enterprise*, in Proceedings of IEA/AIE, London, UK, 1992, Springer-Verlag, pp. 25–34.

[8] W. Gale, K. Church, and D. Yarowsky, *Estimating upper and lower bounds on the performance of word-sense disambiguation programs*, in Proceedings of the 30th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics Morristown, NJ, USA, 1992, pp. 249–256.

[9] S. Golder and B. A. Huberman, *Usage patterns of collaborative tagging systems*, Journal of Information Science 32(2), (2006), pp. 198–208.

[10] M. Guy and E. Tonkin, *Tidying up tags*, 2006.

[11] M. Hepp, *Possible ontologies: How reality constrains the development of relevant ontologies*, IEEE Internet Computing, 11 (2007), pp. 96–102.

[12] A. John and D. Seligmann, *Collaborative tagging and expertise in the enterprise*, in Proceedings of Collaborative Web Tagging Workshop at WWW2006, Edinburgh, UK, 2006.

[13] V. I. Levenshtein, *Binary codes capable of correcting deletions, insertions, and reversals*, Tech. Rep. 8, 1966.

[14] H. Luhn, *The automatic creation of literature abstracts, iBM J*, Res. Develop, 2 (1959), pp. 159–165.

[15] Y. Matsuo and M. Ishizuka, *Keyword Extraction from a Single Document Using Word Co-Occurrence Statistical Information*, INTERNATIONAL JOURNAL ON ARTIFICIAL INTELLIGENCE TOOLS, 13 (2004), pp. 157–170.

[16] P. Mika, *Ontologies are us: A unified model of social networks and semantics*, Web Semantics., 5 (2007), pp. 5–15.

[17] D. R. Millen, J. Feinberg, and B. Kerr, *Dogear: Social bookmarking in the enterprise*, in Proceedings of the SIGCHI conference on Human Factors in computing systems, New York, NY, USA, 2006, ACM, pp. 111–120.

[18] F. Pereira, N. Tishby, and L. Lee, *Distributional clustering of English words*, in Proceedings of the 31st annual meeting on Association for Computational Linguistics, Association for Computational Linguistics Morristown, NJ, USA, 1993, pp. 183–190.

[19] M. Porter, *An algorithm for suffix stripping*, 14 (1980), pp. 130–137.

[20] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA, 1986.

[21] P. Schmitz, *Inducing ontology from flickr tags*, in Proceedings of Collaborative Web Tagging Workshop at WWW2006, Edinburgh, UK, 2006.

[22] L. Specia and E. Motta, *Integrating folksonomies with the semantic web*, in Proceedings of the European Semantic Web Conference (ESWC2007), E. Franconi, M. Kifer, and W. May, eds., vol. 4519 of LNCS, Berlin Heidelberg, Germany, July 2007, Springer-Verlag, pp. 624–639.

[23] J. Surowiecki, *The Wisdom of Crowds*, Anchor, August 2005.

[24] M. Uschold and M. Grüninger, *Ontologies: principles, methods, and applications*, Knowledge Engineering Review, 11 (1996), pp. 93–155.

[25] M. Uschold and R. Jasper, *A framework for understanding and classifying ontology applications*, in Proceedings of the IJCAI99 Workshop on Ontologies and Problem-Solving Methods(KRR5), Stockholm, Sweden, 1999.

[26] M. Uschold, M. King, S. Moralee, and Y. Zorgios, *The enterprise ontology*, Knowledge Engineering Review, 13 (1998), pp. 31–89.

[27] T. Wang and B. Parsia, *Cropcircles: Topology sensitive visualization of owl class hierarchies*, in Proceedings of the International Semantic Web Conference (ISWC2006), I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, eds., vol. 4273 of LNCS, Springer-Verlag, November 2006, pp. 695–708.

Edited by: Dominik Flejter, Tomasz Kaczmarek, Marek Kowalkiewicz
Received: December 1st, 2007
Accepted: January 15th, 2008
Extended version received: June 30th, 2008