Merging of Distributed Topic Maps based on the Subject Identity Measure (SIM) Approach

Lutz Maicher, Hans Friedrich Witschel

University of Leipzig, Department of Information Sciences, Chair of NLP,
Augustusplatz 10-11, 04109 Leipzig, Germany
E-mail: {maicher|witschel}@informatik.uni-leipzig.de

Abstract

The central theoretical criteria of Topic Maps "One Topic for one Subject" leads to serious problems if two distributed Topic Maps are merged: according to existing standards, two Topics will only be merged if the description of their Subject (i.e. their so-called *Subject Identifier* or *Subject Locator*) is exactly identical. On the other hand – from a philosophical point of view – two topics should be merged if they describe the same Subject, i.e. if they are intended to refer to the same thing or idea.

In distributed environments, however, Topic Map authors are not always able to use a common vocabulary: in these cases they will fail to use identical Subject Identifiers/Locators even if they intend to describe the same Subject. Therefore, we propose the SIM (Subject Identity Measure) approach which is based on a statistics using different Topic characteristics. This approach is on the one hand independent of the languages used and on the other hand of the structure in these Topic Maps.

The SIM describes how closely related the Subjects of two distributed Topics are, even if the authors didn't use a common vocabulary. Our algorithm uses as much information as possible in order to support users in decisions about which Topics to merge. If the SIM exceeds a given threshold, this indicates that two Topics describe the same Subject and therefore merging of these Topics will be recommended after a filtering process.

Because Topic Maps are translatable into RDF and OWL the reuse of the SIM in Semantic Web applications should be enforced.

1 Problem

Topic Maps[1] are a powerful tool for Knowledge and Content Management [1]. Derived from their origins as exchangeable indexes for manuals, they are used as powerful access structures for dynamic collections of information resources. Because Topic Maps are a separate layer of metadata on top of the original information resources, they can be seen as the authors' subjective perception of these resources. On the other hand, this separation allows an easy exchange of Topic Maps. If two authors want to share their perception about a specific domain, they merge the according Topic Maps and use the resulting Topic Map conjointly.

Merging is a vital feature of Topic Maps and bases on their central theoretical design criterion. This is called "One Topic for one Subject" (see the Topic Map Reference Model TMRM [6] and the Topic Map Data Model TMDM [14] for further discussion). This means that if two Topics describe the same Subject they must be merged. Although this merging theory is well defined inside the Topic Map standard family, it lacks efficient realisations in practice.

This is due to the fact that Topic Maps are designed to represent subjective perceptions of individual authors, but any merging rule (applied automatically) will have to use some objective criteria. In current standardizations for instance, merging is done by checking whether some Topic properties are exactly identical.

Distributed environments will be an emerging field for the usage of Topic Maps ([2], [10], [12]). Especially in these environments, Topic Map authors will use different vocabulary to describe Subjects. Whenever two Topic authors intend to describe the same Subject, the merging of their Topics might fail if their Subject description is not completely identical.

We propose a similarity measure for Topics which we will call Subject Identity Measure (SIM). This measure describes how closely related the Subjects of two distributed Topics are. The value of the SIM supports humans in their decisions about merging of Topic pairs. Because our approach is language and structure independent, it can be applied to a wide variety of Topic Maps.

Topic Maps are part of the Semantic Web [3] efforts and are translatable into RDF or OWL

[1] To avoid ambiguity all terminology concerning Topic Maps is capitalized.

(discussed in [4], [5], [6]).[2] This enables the reuse of the SIM approach in a variety of Semantic Web applications.

In this paper we are making the following contributions:

- We discuss the merging paradigm of Topic Maps in connection with distributed environments that lack controlled vocabularies (Section 2).
- We introduce a language and structure independent approach for merging of distributed Topic Maps (Section 3).
- We discuss the influence of all parameters introduced by the approach on the basis of a real-life example (Section 4).
- We discuss related work (Section 5) and further research (Section 6).

2 Merging and Topic Maps

The main theoretical design criterion of Topic Maps is called "One Topic for one Subject". In order to understand this criterion, we need to explain the notions of Topic, Subject and their relationship.

A Topic is "a symbol used within a topic map to represent some subject, about which the creator of the topic map wishes to make statements" [14]. A Subject is "anything whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever. In particular, it is anything on which the creator of a topic map chooses to discourse." [14] Shortly, a Topic describes a Subject (which is any possible idea or artefact of discourse) from the perception of the current Topic Map. This implies that within each Topic its Subject must be declared.

Before the Subject of a Topic can be declared, the Topic Map author must be sure of the according Subject. Important philosophical questions arise: What is identifiable? What constitutes the boundaries of a thing in respect to its identity? Can identity evolve in time? Is identity situational or relative? How must properties of a thing change to alter its identity? What about versions and copies?

These questions (discussed in detail in [16], [17]) show the limits of purely computational approaches to merging because they hardly handle indefiniteness, openness and ambiguity. However, we suggest that – despite of the problems mentioned above – a human's decision might be at least supported by computational approaches.

"The process of merging ensures that whenever two topics are known to represent the same subject, they are merged." [14] But how can a Topic declare its Subject? Within the TMDM two (objectively analyzable) means are implemented:

- The *Subject Locator* is used whenever the Subject of the Topic *is* an addressable information resource. In this case, the URI of this resource is used as a Subject Locator.
- Because Subjects can be anything (not only addressable resources) a Topic can declare its Subject with the help of a *Subject Indicator*, too. A Subject Indicator is an information resource which *describes* the Subject. The URI of this information resource is called *Subject Identifier*.

To obtain "One Topic for one Subject", two Topics which have the same Subject Locator or a pair of identical Subject Identifiers have to be merged.[3]

These rules work well if all authors of Topic Maps have made agreements about a centralised conceptualisation of the represented knowledge. These agreements are called *Published Subject*

[2] The Topic Map browser "Omnigator" from the Topic Map vendor Ontopia (http://www.ontopia.net) allows the export of Topic Maps into RDF and the import of RDF data.

[3] In [14] additional equality rules are defined. Two Topic items must be merged if they have the same Source Locator (ID) or the Subject Identifier of the first Topic is the Source Locator of the second Topic. In XTM 1.0 topics are defined as identical if they have an equal Basename in the same Scope. See [19], §10 for the criticism of the latter approach.

Indicators (PSI) [17]. These PSIs are published (but not necessarily public) descriptions of Subjects which should be reused by as much Topic Map authors as possible in order to obtain a broad interoperability of their Topic Maps. Examples in the literature which discuss the merging of distributed Topic Maps (or Topic Maps and RDF documents) exclusively use PSIs (see [7], [8]).

However, in distributed environments (which should be preferred for KM applications [7], [10], [23]), with a high autonomy of the clusters, the mechanism of PSIs has its shortcomings. A PSI will only be used if it is visible to a Topic Map author. If it isn't, authors will tend to create and use their own private Subject Indicators.

But if no PSIs are used, merging of Topic Maps becomes impossible because there will probably be no common Subject Indentifiers/Locators. And this might happen even if the Topic Map authors made assertions about the same Subjects in their private Topic Maps: If the distributed authors used different Subject Indicators to indicate the same Subject, the regarding Topics, which should theoretically be merged, rest apart.

But "Merging beyond the minimal rules [defined in the TMDM] is freely allowed. Most commonly, this will be done by inferring the subject of the topics from their characteristics." [14] We will accept this recommendation.

Therefore, we propose a Subject Identity Measure (SIM). If this measure is 1 the regarding Topics definitely represent the same Subject (according to the rules defined in the TMDM). If the measure is 0, the regarding Topics definitely represent different Subjects. All values between 0 and 1 support a human being to decide whether two Topics represent the same Subject. In cases where recall is more important than precision, all Topics that have a SIM which is higher than a certain threshold will automatically be proposed for merging.

3 Overview of the SIM Approach

The goal of the SIM approach is to be SIMple and of high quality. For the calculation of the SIM we only use the data inside each Topic. As we mentioned above, we don't use structural information (types, associations etc.). Therefore, we can regard our approach as a simple and lightweight solution which is an ideal benchmark (or baseline). Each solution which is more advanced (and more computationally expensive) can be compared to the results given by the SIM: in regard to precision, recall, F-value and performance.

Whenever two Topic-Maps meet, our approach performs the following steps:

1. *Calculation.* The SIMs for Topicnames[4], Occurrences[5] and Subject Indicators for each pair of Topics must be calculated.
2. *Filtering.* According to different thresholds and coefficients, the overall SIM will be calculated. For each Topic, a suitable counterpart in the other Topic Map will be chosen: the one that has the greatest SIM (but only if the SIM is greater than 0).

In the following, these two steps are discussed in detail.

Calculation

We want to find simple, language and structure independent similarity measures for each data item occurring in a Topic. According to the TMDM we have to handle URIs (for Subject Indicators, VariantNames and OccurrenceLocators) and strings (all TopicNames and OccurrenceData).

For a pair of two strings *(S1,S2)* we calculate a similarity measure as follows:

[4] Topics may have different Names (with different Types and different Scopes). "A topic name is a name for a topic, consisting of the base form […] and variants of that base form". [14]

[5] An occurrence is a representation of relationship between a subject and an information resource" [14]. In most cases, occurrences are something like example sentences (in which the subjects *occurs*). Occurrences are either Strings or URIs

1. Remove from *S1* and *S2* all special characters and numbers.
2. Remove from *S1* and *S2* all words that have less than a fixed number (e.g. 4) of characters.
3. Let $|S1| = m$ and $|S2| = n$ and let $m<n$. For each token *t* in *S1*, decide if this token occurs in *S2*, too. If it does, increment c (starting from 0) by 1.
4. The similarity measure *s(S1,S2)* is:

$$s(S1,S2) = \frac{c}{m}$$

So *s(S1,S2)* tells how many of the possible word matches between *S1* and *S2* have been found (note that there can only be a maximum of *m* matches!).

This method is independent of the language used in the strings and the context where the string occurs.

For URIs *U1* and *U2*, we propose the following approach to get their similarity *c(U1,U2)*. According to the W3C, an URI identifies a resource, which might have a representation (e.g. a web page). Normally, URIs consist of the following parts which should be interpreted separately [11]:

- a *scheme* (e.g. "http"), which declares how the URI should be interpreted and which protocols should be used to get a representation of the resource,
- an *authority* (e.g. "www.km.org") which owns the URI,
- the *path* to the resource (e.g. "/style/test.htm") to distinguish the authority's resources unambiguously and
- a *fragment* of the resource (e.g. "#mypart").

Although we didn't implement these comparisons, we will sketch some thoughts on them. Basically, one has to decide whether the URIs of the resources, their representations or both should be looked at. For simplicity we suggest only a URI comparison.

According to the TMDM, *c(U1,U2)* must be 1 if *U1* and *U2* are identical byte by byte. Calculating *c(U1,U2)* will be checking if any parts of the URI are identical. For example, one could assume that if two URIs refer to the same resource (identical schema, authority and path) and only the fragments differ, the *c(U1,U2)* should be greater than 0. But we have to bear in mind that the authority chose different URIs intentionally. Therefore we can assume that the URIs perhaps indicate similar Subjects but with a high probability not identical ones. While for the context of Subject Identifiers these thoughts indicate a *c(U1,U2)* near zero, for Occurrences the similarity is important and indicates a *c(U1,U2)* near to 1 (if two Topics have very similar Occurrences, they could be identical). This tells us that this measure can't be context independent. Because we have no empirical results, we won't take this discussion into more detail.

For the following description of the calculation algorithm, we only use *s(S1,S2)* as introduced above, i.e. our current approach handles string data only and doesn't consider any URIs.

The algorithm inspects each possible pair of Topics *(T1,T2)* where *T1* must belong to Topic Map 1 and *T2* must belong to Topic Map 2. For each pair a SIM.Names and a SIM.Occurrences is calculated as follows:

1. ***"Fillet" each Topic for Names.*** Take all property values from the property "value" of all Topic Name Items and Variant Items of T1 and store them in a set *Nam1*.[6] To get *Nam2* do the same for T2.
2. ***"Fillet" each Topic for Occurrences.*** Take all property values from the property "value" of all Occurrence Items of T1 and store them in a set Occ1. The same for T2.
3. ***Calculate SIM.Names.*** If $|Nam1| < |Nam2|$,[7] and $|Nam1|=m$ then:

$$SIM.Names = \frac{1}{|m|} \sum_{n1 \in Nam1} \max_{n2 \in Nam2} s(n1,n2)$$

This means that we calculate the "best

[6] To lay aside the terminology of TMDM: Get all Strings from all Base- and VariantNames of the current Topic regardless of their Types and Scopes.

[7] This decision follows the principle of the "weakest link".

match" for each element of *Nam1* and then average over all these numbers.

4. ***Calculate SIM.Occurrences*.** Do the same for *Occ1* and *Occ2*.

Using SIM.Names and SIM.Occurrences, we can now determine which Pairs of the two Topic Maps should be merged.

Filtering

With the help of *SIM.Names* and *SIM.Occurrences,* SIM(T1,T2) for each pair of Topics (T1,T2) is calculated as follows:

$$SIM = \lambda SIM.Names + (1-\lambda) SIM.Occurrences$$

But: SIM = 0 if SIM.Names < t_{Name} or SIM.Occurrences < t_{Occ} or SIM < t

SIM depends on three parameters:

- t_{Name}. SIM.Names must exceed this threshold to avoid overrating of SIM.Occurrences if SIM.Names is small.
- t_{Occ}. SIM.Occurrences must exceed this threshold to avoid overrating of SIM.Names if SIM.Occurrences is small.
- λ. This parameter indicates whether Names or Occurrences are more important. If $\lambda=1$, only Names are of interest, if $\lambda=0$ we only look at Occurrences.
- t. The overall SIM must exceed this threshold to avoid a great number of false positives (i.e. to improve precision).

With the help of *SIM(T1,T2)* for each pair we extract merging candidates. In our case, we use the application-specific constraint that inside a Topic Map each Subject is represented by only one Topic. This means that for each Topic in one Topic Map we can find at most one Topic to merge in the opposite Topic Map.

We iterate over the smaller Topic Map. For each Topic *T1*, we choose a merging candidate *T2* such that *T2* is the maximum over all *SIM(T1,Ti).* If *SIM(T1,T2)* is *0,* the Topic *T1* remains without a merging candidate.

We assume that the parameters must be varied according to the Topic Maps which are merged. It is not even clear if all of them are really necessary.

In the following section, we discuss the optimal choice of these parameters for a specific example and try to determine the really relevant ones.

Remarks

To get a very simple approach (which should be treated as a reference for more advanced approaches) we decided not to make use of the following information:

- *Associations*. The associations a Topic is involved in might be very good indicators for its Subject.
- *The representation of the referenced information resources*. The content of these representations referenced by an URI might be a good indicator for Subjects. For a light weighted approach we proposed to interpret solely the URIs.
- *Types and Scopes*. Types and Scopes indicate the Subject of a given Topic, too (esp. for Scopes see [12]). Because both are sets of Topics, we have to pay attention to not running into recursivity problems.

4 Assessment of Matching Quality

For evaluation purposes, we needed Topic Maps which describe fully qualified Subjects with uncontrolled vocabularies in a slightly subjective manner. For this, we decided to use the online catalogues of two different libraries. The first is the German Library ("Deutsche Bücherei, DDB", http://www.ddb.de) and the second is a compound catalogue of a network of German libraries ("GBV", http://gso.gbv.de/).

We extracted from these catalogues all entries which represent publications of Springer in 1997. These are approximately 1800 entries in the catalogue of the DDB and approximately 2700 entries in GBV.
For each entry, we put aside the ISBN or ISSN as an objective criterion helping to decide whether two publications in the two catalogues were

indeed identical. Then we measured the results of our SIM against this objective criterion.

We observed that the description and the kind of available information in the two catalogues differed significantly, i.e. that the two Topic Maps we built from them were sufficiently different to provide a good test for our SIM measure.

From these datasets, we automatically created Topic Maps where each dataset (i.e. each description of a publication) is represented by a Topic. Some properties like "title" are made Basenames of these Topics, the other properties (e.g. "keywords" or "editor") are treated as typed Occurrences of these Topics. We obtained a Topic Map without Associations.

To assess the quality of our results obtained with the SIM approach, we use three quality metrics. Assume that G is the set of all pairs found by the System and I is the set of all pairs which are identical (confirmed by identical ISBN/ISSN):

1. *Precision* tells how many of the merging candidates proposed by SIM are really identical:
$$P = \frac{|G \cap I|}{|G|}$$

2. *Recall* tells how many of the existing identical pairs were found by SIM:
$$R = \frac{|G \cap I|}{|I|}$$

3. *F(β)-Value:* A combination of P and R that yields high values only if both P and R are high. The formula
$$F(\beta) = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$
allows to weight P and R differently (see [13]). For example, if β=2, then R is twice as important as P. We chose F(2) as our measure because recall is much more important with merging Topic Maps: finding merging candidates that SIM failed to find is much more tedious than eliminating some false positives.

We introduced four parameters which might influence the quality of our approach. At first, we eliminate all parameters (λ*=0.5,* t_{Occ}*=0* and t_{Name}*=0*) with the exception of *t*. As an example set, we randomly picked approximately 300 Topics from each catalogue.[8] We obtained around 90.000 possible Topic pairs where 25 are positive matches based on ISBN equality. Precision, recall and F(2) for this example are presented in **Figure 1**.

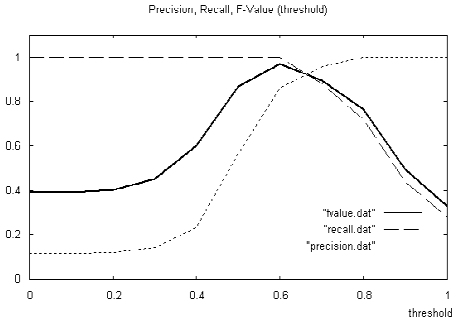

Figure 1 Precision, Recall, F-Value (*t*), [λ*=0.5,* t_{Occ}*=0.0,* t_{Name}*=0.0*]

As the threshold increases, recall remains *1* until the global maximum of *F(2)=0.96* is reached with *t=0.6*. After this maximum (we obtained all 25 true positive and 4 false positives, too), the recall decreases rapidly while precision reaches its maximum.
To confirm these results, we randomly created four example sets of a similar size and calculated the behaviour of *F(2)* as a function of *t*. **Figure 2** shows the results. With *t=0.5,* we obtain an average *F(2)=0.84* and with *t=0.6* we obtain an average *F(2)=0.89* over all five sets. We assume that with *t=0.6* and elimination of all other parameters similar results will be obtained for all possible sets of our testbed.

[8] Our approach favours large Topic Maps. Therefore we decided to take only a fraction of our testbed for evaluation purposes as a sort of "worst case scenario".

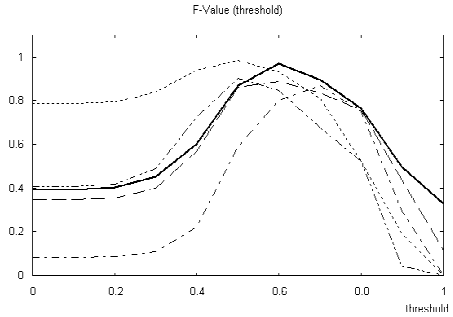

Figure 2 F-Value (*t*), [λ*=0.5,* t_{Occ}*=0.0,* t_{Name}*=0.0*] for all five example sets

Now we have to discuss the influence of the different parameters on our result. At first, we want to examine whether t_{Name} and t_{Occ} might help to eliminate the significant number of false positives without loss of true positives.

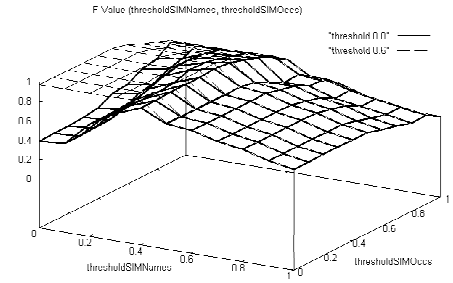

Figure 3 F-Value (t_{Occ}, t_{Name}), [λ*=0.5*] for *t=0.0* and *t=0.6*

We see (**Figure 3***) that* the combination of *t*, t_{Names} and t_{Occ} leads to significantly better results. For *t=0.6* we obtain a global maximum *F(2)=0.98* for t_{Names}*=0.4* and t_{Occ}*=0.5* with 25 true positives and only 2 false positives). If *t* is eliminated *(t=0.0)* we obtain a global maximum *F(2)=0.93* for t_{Names}*=0.4* and t_{Occ}*=0.5* but 9(!) false positives for 25 true positives. We notice that in the vicinity of the maximum t_{Names} and t_{Occ} are smaller than *t*. Because we could observe a quite similar behaviour for all five example sets, we derive the assumption that a combination of *t*, t_{Names} and t_{Occ} is fruitful.

The last parameter to be discussed is λ. Because in our example the Topicnames (i.e. the titles of books) might be more discriminating than the Occurrences (i.e. the keywords, authors etc. describing the books), we expect increasing F-values for decreasing λ. In **Figure 4** we examined the behaviour of F(2) in respect to λ in four scenarios. The scenarios with *t=0.0* simulate the elimination of this parameter, while t=0.6 simulates the behaviour in the global maximum. The scenarios with t_{Names}*=0* and t_{Occ}*=0* simulate the elimination of these parameters while t_{Names}*=0.4* and t_{Occ}*=0.5* simulate the behaviour in the global maximum.

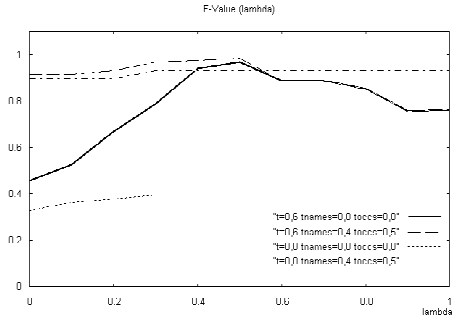

Figure 4 F-Value (λ) for *(t=0.0;* t_{Name}*=0.0;* t_{Occ}*=0.0), (t=0.0;* t_{Name}*=0.4;* t_{Occ}*=0.5), (t=0.6;* t_{Name}*=0.0;* t_{Occ}*=0.0), (t=0.6;* t_{Name}*=0.4;* t_{Occ}*=0.5)*

We obtained some surprising results. For the given example set, in all four scenarios the maximum is reached for λ*=0.5* which indicates that varying λ doesn't improve the results. To test this hypothesis, we repeated this test for all five example sets with *t=0,* t_{Names}*=0* and t_{Occ}*=0* (which was the scenario with the largest variance).

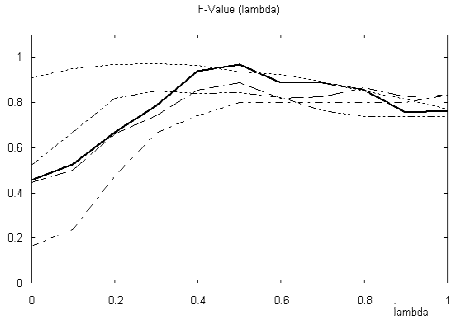

Figure 5 F-Value (λ), [*t=0.6;* t_{Name}*=0.0;* t_{Occ}*=0.0*] for all five example sets

The results shown in **Figure 5** strengthen our hypothesis. In the average case, we obtain very good results with the elimination of λ. This confirms our assumption that the usage of λ doesn't improve the overall quality of our approach.

Our previous results recommend the following values of the parameters: $t=0.6$, $t_{Names}=0.4$ and $t_{Occ}=0.5$ which yields a very good F-Value of 0.98. In other scenarios, however, the optimum might be reached with other thresholds. Generally, we recommend the usage of low thresholds so as to obtain high recall (as mentioned above, the elimination of some false positives is not as tedious as finding new candidates for merging).

Summarising, we sketch the following results:

- Our approach yields good results for both recall *and* precision.
- We propose the usage of t, t_{Names} and t_{Occ}
- If t_{Names} and t_{Occ} are chosen carefully, the usage of t isn't necessary.
- λ doesn't augment the overall quality of our approach and can be eliminated.

These results are derived from measurements based on our library example. But we must bear in mind that for other example sets, especially for Topics with few strings, our approach must be modified and advanced for future challenges. This is discussed in the section "Further Research".

5 Related Work

Basically, our problem resembles schema and ontology matching, in particular the special case of *instance based matching*. Therefore, we refer (for a detailed overview about matching techniques) to [20]. According to the classification given there, our approach is an individual matcher which is instance/contents-based on an element level and which deals with (very light weighted) linguistic features.

We want to pick out only one promising approach. The *similarity flooding* approach proposed by Melnik et al. [21] might be very interesting, because it is language and structure independent in a radical way. Because this approach even disregards the semantic of the Topic Map Items, it will be interesting if RDF statements and Topic Maps should be merged. We foresee complexity problems because the computational costs of the approach are higher than with our approach. It should be tested if similarity flooding can be extended to Topic Maps with several hundred Topics.

Most of the approaches in [20] avail or propose the usage of thesauri, taxonomies or ontologies to derive semantical similarity between strings. The usage of these tools should augment the matching quality. The price you have to pay for this is a drastic loss of flexibility: the approach becomes language dependent. Alongside with our results, the findings in [21] promise quite powerful solutions without the usage of these tools.

6 Further Research

The proposed SIM can be used to build systems which support users in (peer-to-peer) merging scenarios like Semantic Web applications.
We didn't discuss the problem of Subject Indicator harmonisation. Even if the user decided to merge two Topics, their Subject Indicators must be harmonised. Only if the Subject Indicators are identical, these Topics will be merged according to the TMDM. This is an issue of harmonising the given Subject Indicators of the Topics into one vocabulary.

If such systems exist, *relevance feedback* can be used to enhance recall, precision and performance. For example, in a first step, one might calculate SIM only for pairs of Topics which are used as Types. For this task, we propose the usage of the approaches introduced in the previous section (esp. [21], [22]). When the merging of Types (supervised by the user) has been completed, this knowledge can be used in subsequent steps: Topics with the same type are more likely to receive high SIM values. In these cases our approach remains structure independent because the assumptions about the structure are not inferred from the "outside" of the algorithm.

The same applies to Associations between Topics: If the set of neighbours of one Topic is similar to the set of neighbours of a Topic in another Topic

Map, this can be interpreted as an indication of a high SIM. This is, of course, recursive because Topic Maps are networks: the similarity of the neighbourhood sets is based on the SIM measure, too. The similarity flooding proposed in [21] shows how this recursivity can be avoided.

The approach should be examined with more different example sets to test our hypothesis. In addition, it might be interesting if t_{Names} and t_{Occ} can be harmonised to one threshold which is valid for SIM.Names and SIM.Occurrences. Furthermore, examples should be examined that deal with Locators to extend the SIM approach with respect to URIs.

Conclusion

We showed that the existing merging rules of Topic Maps have their limitations in distributed scenarios where Topic Map authors might not share a common vocabulary to declare Subjects. A merging approach on top of these rules is needed.

We proposed the Subject Identity Measure (SIM) which describes how closely related the Subjects of two distributed Topics are, even if the authors didn't use a common vocabulary. The SIM builds a bridge between the intrinsic subjectivity of real Topic Maps and the intrinsic objectivity of the declaration of Subjects.

For our example, the SIM approach yields good results for both recall *and* precision. We consider the introduced approach as a benchmark: each advancement should be tested against the quality of the SIM approach.

When discussing the influence of some free parameters, we suggested some simplifications by elimination of parameters that didn't decrease the quality of the approach in the examined example.

We have seen that the proposed SIM has some limitations which must be eliminated in further research. In doing so, a lot of work contributed by schema matching approaches should be considered.

References

[1] G. Heyer, L. Maicher: *"Persönliche und gemeinschaftliche Wissensräume. Erfüllen Topic-Maps die technologischen Anforderungen?"* In: K.-P. Fähnrich, H. Herre (Eds.): Content- und Wissensmanagement. Beiträge auf den LIT'03. Leipzig (2003).

[2] K. Ahmed: *"TMShare – Topic Map Fragment Exchange In a Peer-to-Peer-Application."* In: Proceedings of XML Europe 2003, (2003).

[3] T. Berners-Lee, J. Hendler O. Lassila: *"The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities."* In: Scientific American. May (2001), pp.34-43.

[4] S. Pepper: *"Ten Theses on Topic Maps and RDF."* Available at: http://www.ontopia.net/topicmaps/materials/rdf.html

[5] L. M. Garshol: *"Living with topic maps and RDF. Topic maps, RDF, DAML OIL, OWL, TMLC."* Available at: http://www.ontopia.net/topicmaps/materials/tmrdf.html

[6] ISO/IEC JTC 1/SC34: *"Topic Maps – Reference Model. Editor's Draft, Revision 3.1. 01.12..2003."* Available at: http://www.isotopicmaps.org/TMRM/TMRM-latest-clean.html

[7] P. Ciancarini, M. Pirruccio, F. Vitali, R. Gentilucci, V. Presutti: *"Metadata on the Web. On the integration of RDF and Topic Maps."* In: Proceedings of "Extreme Markup Languages 2003", Montreal, 2003.

[8] G. O. Grønmo: *"Automagic Topic Maps."* Available at: http://www.ontopia.net/topicmaps/materials/automagic.html

[9] R. Cuel: *"A New Methodology for Distributed Knowledge Management Analysis."* Proceedings of I-KNOW '03, Graz, (2003), 531-537.

[10] T. Schwotzer: *"Modelling Distributed Knowledge Management Systems with*

Topic Maps." Proceedings of I-KNOW '04, Graz, (2004), to appear.

[11] I. Jacobs (eds.): *"Architecture of the World Wide Web, First Edition. W3C Working Draft 5 July 2004.*" Available at: http://www.w3.org/TR/webarch

[12] A. Sigel: *„kPeer as a Context-Aware Topic Map P2P Application for the Distributed Integration of Knowledge"*. Submitted to MRC 2004. Available at: http://kpeer.wim.uni-koeln.de/~sigel/

[13] E. Riloff, W. Lehnert: *„Information extraction as a basis for high-precision text classification."* In: ACM Transaction on Information Systems, 12 (3), pp. 296-333 (1994).

[14] ISO/IEC JTC 1/SC 34: *"ISO/IEC 13250. Topic Maps – Part 2: Data Model."* Latest version available at: http://www.isotopicmaps.org/sam/

[15] TopicMaps.Org Authoring Group: *"XML Topic Maps (XTM) 1.0."* Available at http://www.topicmaps.org/xtm/1.0/

[16] W. Kent: *"The unsolvable identity problem."* In: Proceedings of "Extreme Markup Languages 2003", Montreal, (2003).

[17] W. Kent: *"Data and reality. Basic Assumptions in Data Processing Reconsidered."* North-Holland Publishing, Amsterdam, New York, Oxford (1978).

[18] OASIS: *"Published Subjects: Introduction and Basic Requirements."* Available at: http://www.oasis-open.org/committees/download.php/3050/

[19] E. Freese: *"So why aren't Topic Maps ruling the world?"*, Proceedings of „Extreme Markup Languages 2002", Montral (2002).

[20] E. Rahm, P. A. Bernstein: "*On Matching Schemas Automatically."* Technical Report MSR-TR-2001-17. Available at: http://www.research.microsoft.com/pubs/

[21] S. Melnik, H. Garcia-Molina, E. Rahm: "*Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching."* In: Proceedings of 18th International Conference on Data Engineering (ICDE'02), San Jose, California, (2002).

[22] S. Castano, A. Ferrara, S. Montanelli, G. Racca: *"Matching techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions.*" In: Proceedings of International Conference on Coding and Computing (ITCC04), IEEE Computer Society, Las Vegas, (2004).

[23] K. Böhm, L. Maicher, H.-F. Witschel, A. Carradori: *"Moving Topic Maps to Mainstream - Integration of Topic Map Generation in the User's Working Environment."* Proceedings of I-KNOW '04, Graz, (2004), to appear.

[24] S. Pepper: *"Towards a General Theory of Scope."* Proceedings of „Extreme Markup Languages 2001", Montral (2001).

[25] K. Böhm, G. Heyer, U. Quasthoff, C. Wolff: *"Topic Map Generation Using Text Mining."* J.UCS - Journal of Universal Com-puter Science (Springer), Volume 8, Issue 6, S. 623-633.