An Empirical Study of Instance-Based Ontology Matching

Antoine Isaac[1,2], Lourens van der Meij[1,2], Stefan Schlobach[1], and Shenghui Wang[1,2]

[1] Vrije Universiteit Amsterdam
[2] Koninklijke Bibliotheek, Den Haag
{aisaac,lourens,schlobac,swang}@few.vu.nl

Abstract. Instance-based ontology mapping is a promising family of solutions to a class of ontology alignment problems. It crucially depends on measuring the similarity between sets of annotated instances. In this paper we study how the choice of co-occurrence measures affects the performance of instance-based mapping.

To this end, we have implemented a number of different statistical co-occurrence measures. We have prepared an extensive test case using vocabularies of thousands of terms, millions of instances, and hundreds of thousands of co-annotated items. We have obtained a human Gold Standard judgement for part of the mapping-space. We then study how the different co-occurrence measures and a number of algorithmic variations perform on our benchmark dataset as compared against the Gold Standard.

Our systematic study shows excellent results of instance-based matching in general, where the more simple measures often outperform more sophisticated statistical co-occurrence measures.

1 Introduction

Dating as far back as the problems of record and database schema integration, studied for well over 40 years now, the semantic heterogeneity problem is probably the single-most urgent problem to be solved to realize a web-scale Semantic Web. A huge number of ontologies is now available.[1] This makes automatic ontology mapping,[2] as the anticipated solution to semantic heterogeneity, is therefore a research issue of paramount importance. To address it the Semantic Web community has invested significant efforts over the past few years. This has led to the development of a plethora of high-quality matching software, whose potential has been proven in specific applications in a variety of domains.

Ontology mapping techniques are commonly divided into 4 broad categories [1]: lexical (detecting similarities between labels of concepts), structural (using

[1] Swoogle has indexed over 10,000 of them, cf http://swoogle.umbc.edu/.

[2] *Ontology mapping* is the task of determining relations such as equivalence or subsumption between concepts of two separate ontologies.

K. Aberer et al. (Eds.): ISWC/ASWC 2007, LNCS 4825, pp. 253–266, 2007.

the structure of the ontologies), based on background knowledge, and instance-based mapping (using classified instance data). Among these, there are surprisingly few systematic studies of instance-based ontology mapping, *i.e.* the construction of links between concepts based on the co-occurrence of instances. In Chapter 6.2 of [1] a number of systems are discussed which make use of extensional information. However, to the best of our knowledge there has been no systematic evaluation yet over the use of different similarity measures for instance-based mapping, and this paper attempts to close this gap.

The basic idea of instance-based mapping is that the more significant the overlap of common instances of two concepts is, the more related these concepts are. The difficult question is how to define the notion of significance for such *extension* overlap. Previous investigations on instance-based mapping [2,3] have shown that there are some crucial decisions to be made with this respect. We propose a systematic approach considering the following dimensions:

- **Measures:** the most simple idea is to calculate the common factor of two concepts C and D as the proportion of jointly annotated books over the sum of books annotated by C and D, as done by the Jaccard measure. In statistics and Information Theory a number of other measures have been developed, such as Pointwise Mutual Information, Information Gain or Log-likelihood ratio.
- **Thresholds:** often the above mentioned measures are vulnerable for data-sparseness: if there are too few instances, the common factor measure ranks mappings high when the two concepts involved can only be found in one single book's annotation. The solution to dealing with this issue is to consider thresholds in the measures.
- **Hierarchy:** following the semantics of ontologies we can use the hierarchy, *i.e.* including the instances of descendants in the extension of a concept.

In this paper we will study the effect of these choices on a critical application in which it is considered to combine two thesauri. We implemented a system that calculates ranked lists of mappings according to 5 measures and different thresholds. It also allows us to include instances from a concept's descendants into its extension. We evaluated the resulting mappings against a Gold Standard built manually.

Based on this case-study we will answer the following research questions

1. Is instance-based mapping a reliable technology to be applied in practical, possibly critical applications?
2. Which combination of measures, thresholds and information inclusion works best, possibly depending on circumstances such as whether precision or recall is considered more important?

The first question can be answered easily: our results show an excellent level of quality. The second, more technical question will be answered in the course of the paper.

It is worth emphasising that we make the non-trivial assumption that doubly annotated instances exist. Furthermore, note that we evaluate the quality of

the similarity measures rather than compare performances of (existing or new) mapping systems. For a discussion on the use of different measures and methods for possible application scenarios for mappings we refer to [4].

The paper is structured as follows. In Section 2 we introduce our application. In Section 3 we describe the methodology of our mapper, including the different measures and parameters. In the remaining sections we will describe our experiments and the results before Section 6 sums up our findings, and presents perspectives on future work.

2 Use Case Scenario

The National Library of the Netherlands[3] maintains a large number of collections. Two of them are the *Deposit Collection*, containing all the Dutch printed publications (one million items), and the *Scientific Collection*, with about 1.4 million books mainly about the history, language and culture of the Netherlands. Each collection is described according to its own indexing system. On the one hand, the Scientific Collection is described using the GTT, a huge vocabulary containing 35,000 general terms ranging from *Wolkenkrabbers* (Skyscrapers) to *Verzorging* (Care). On the other hand, the books contained in the Deposit Collection are mainly indexed against the *Brinkman thesaurus*, containing a large set of headings (more than 5,000) that are expected to serve as global subjects of books. Both thesauri have similar coverage but differ in granularity. Also, for each concept, they provide the usual lexical and semantic information found in thesauri: synonyms and notes, broader and related concepts, etc.

The co-existence of these different systems, even if historically and practically justified, is not satisfactory from the point of view of interoperability. The KB is therefore investigating ways to combine the two thesauri, trying to enhance integration while ensuring compatibility with legacy data of both systems. For this reason, mapping GTT concepts with Brinkman concepts is crucially needed.

Finally, it is important to mention that around 250,000 books are common to the depot and scientific collections, and have therefore been manually annotated with both GTT and Brinkman vocabularies. This makes the KB use case especially suitable for studying instance-based mapping techniques.

3 A Framework for Instance-Based Mapping

We will now describe our formal framework for instance-based mappings, slightly adapting the one presented in [5]. Given two ontologies $\mathcal{S}$ (for source) and $\mathcal{T}$ (target) we see a mapping as a triple (S, T, R), where R is a relation between concepts $S \in \mathcal{S}$ and $T \in \mathcal{T}$. Often, the relation R is taken from the set $\{\equiv, \sqsubseteq, \sqcap, \perp\}$, resp. for equivalence, subsumption, overlap and disjointness. In an application about thesauri [6], relations similar to *broader than*, *narrower than*, and even the *related to* relation might also be considered.

[3] Koninklijke Bibliotheek (KB), `http://www.kb.nl`

In instance-based mapping semantic relations between concepts of two ontologies are determined based on the overlap of their instance sets. This is a very natural approach, as in most ontology formalisms the semantics of the relations between concepts is defined via the set of their instances. The idea for mapping is then simply that the higher the ratio of co-occurring instances for two concepts, the more related they are.

As instance-based mapping is closely depending on the meaning of a concept in an ontology formalism, **different ways of interpreting concepts** have to be taken into account. The most prominent question is whether a concept is interpreted as the collection of instances annotated by itself alone, or whether the instances of its descendants in the hierarchy also belong to its extension.

Unfortunately, in the real world we also have to deal with incorrectly annotated instances, data spareness and ambiguous concepts, so that basic statistical measures of co-occurrence, such as the Jaccard measure, might be inappropriate if applied in a naive way.

We deal with this problem in two ways: first **we use other measures for calculating relatedness** of sets based on their elements, such as Pointwise Mutual Information, Information Gain or Log-Likelihood ratio, which have been developed in information theory and statistics. Finally, we **consider statistical thresholds** which explicitly exclude statistically unreliable information.

This analysis immediately suggests a systematic study of different instance-based mapping paradigms according to three dimensions:

Measures — Hierarchy — Thresholds

We will use these in the following sections to answer the research questions we outlined in the Introduction, based on a set of systematic experiments. First however, let us briefly fix some technical terms we use later in the paper.

3.1 Measures

We use similarity measures to order pairs of proposed mappings according to the strength of their relatedness, and in our experiments we assess the ranking rather than the objective values. Therefore, we do not need any special normalisation of measures, nor require them to be within the 0-1 interval.

In the following we will call the set of instances annotated by a concept C its extension, and abbreviate by C^i. As usual the cardinality of a set S is denoted by $|S|$.

Jaccard Measures. The first candidates are functions that measure the fraction of instances annotated by both concepts relative to the set of instances annotated by either one of the concepts.

Jaccard. The first measure

$$JC(S,T) = \frac{|S^i \cap T^i|}{|S^i \cup T^i|} \tag{1}$$

is known as the Jaccard measure. If there is a perfect correlation between two concepts, the measure will have a value of 1; if there is no co-occurrence, the measure will be 0. An evident problem with this measure is that it does not distinguish between two matches $(S,T),(S',T')$ where the first tuple co-occurs in 100 instances while the second is based on a single book containing (T'_1,T'_2) in both cases with no other occurrences of both concepts. Yet, a mapping based on one instance gives intuitively less evidence for equivalence than the case based on 100 different books.

Corrected Jaccard. To correct this, we define *corrected Jaccard* with the goal of assigning a smaller score to less frequently co-occurring annotations. We (relatively arbitrary) choose a factor of 0.8 so that evidence based on one co-occurring instance is weighed as much as mapping two concepts would get when a large number of concepts have 20% in their intersection.

$$JCcorr(S,T) = \frac{\sqrt{|S^i \cap T^i| \times (|S^i \cap T^i| - 0.8)}}{|S^i \cup T^i|} \tag{2}$$

Standard Information-Theory Measures. Similarity measures for concepts based on annotations is not new, and often standard statistical measures have been applied to extract semantics from natural language texts based on co-occurrence of terms. As the problem is closely related to mapping concepts, we consider three of those measures: Pointwise Mutual Information, Log-Likelihood ratio and Information Gain.

Pointwise Mutual Information. Pointwise Mutual Information measures the reduction of uncertainty that the annotation of one concept yields for the annotation with the other. For mapping we use co-occurrence counts to estimate probabilities:

$$PMI(S,T) = log_2 \frac{|S^i \cap T^i| \times N}{|S^i| \times |T^i|} \tag{3}$$

where N is the number of annotated instances.

Log Likelihood ratio. In the context of word co-occurrence in corpora it was noticed that PMI is inadequate to deal with sparse data [7]. Data sparseness is also a problem in our case, because the set of annotated objects is often small as compared to the size of the ontologies.

For the likelihood ratio, we compare the hypothesis that p_1 is the maximum likelihood estimate of the probability $P(i_2|i_1) = \frac{k_1}{n_1}$ and that p_2 is the maximum likelihood estimate of the probability $P(i_2|\neg i_1) = \frac{k_2}{n_2}$, with the hypothesis that $p_1 = p_2 = P(i_2) = \frac{k_1+k_2}{n_1+n_2}$, which is just the maximum likelihood estimate of the probability of i_2.

In order to scale this ratio to make comparison possible, we use the log-likelihood form $-2\log\lambda$. Thus, for our particular situation, we compute:

$$-2[\log L(p_0,k_1,n_1) + \log L(p_0,k_2,n_2) - \log L(p_1,k_1,n_1) - \log L(p_2,k_2,n_2)]$$

where $\log L(p,k,n) = k \log p + (n-k)\log(1-p)$
and $k_1 = |S^i \cap T^i|$ $k_2 = |S^i| - |S^i \cap T^i|$ $n_1 = |T^i|$
and $n_2 = N - |T^i|$ $p_1 = k_1/n_1$ $p_2 = k_2/n_2$ $p_0 = |S^i|/N$.

Information Gain. Information gain is the difference in entropy, i.e. the amount of information we can gain about a hypothesis by observing, and is used in decision trees learning to determine the attribute that distinguishes best between positive an negative examples.

In ontology mapping the analogy is the following: Information Gain describes the in- or decrease of the difficulty of assigning a concept to an instance if it has already been annotated with a concept from the other ontology. Formally, the entropy of assigning a concept T to an instance i can be estimated by $e_1 = -\frac{|T^i|}{N} \times log2(\frac{|T^i|}{N})$, where N is again the number of instances. After assigning a concept S from the source ontology the entropy $e_2 = -\frac{|S^i \cap T^i|}{|S^i|} \times log2\frac{|S^i \cap T^i|}{|S^i|}$. The information gain is then $IG = e_1 - e_2$.

For Information Gain the order of source and target are of crucial importance. For mapping targeting equivalence and relatedness, however, we do not have to take symmetry information into account. The version used in our experiments is a combination of two IG measures: $IGB(S,T) = max\{IG(S,T), IG(T,S)\}$.

3.2 Enforcing Thresholds to Guarantee Statistical Relevance

Both Log-likelihood and Information Gain take the number of instances of a concept into account to ensure statistical viability of its results. An alternative approach is to set a threshold for discarding computation of measures if the extension of one of the concepts is too small. The aim of this study is not to find the ideal threshold for statistical relevance, because this will probably too strongly depend on the collection of instances. However, we want to empirically show that there is a difference between using a threshold, and not using a threshold. Therefore, we only consider values 1 and 10 for cut-off, and denote a measure M with cut-off as M_{10}.

3.3 Hierarchical Information

For all the measures we previously defined, we used as the extension of a concept C the set C^i of its direct instances, *i.e.* the set of books explicitly annotated with it. However, semantically, this is not the only option, as one could also take more information from the ontology into account. Especially, a *broader than* relation could imply that the instances of the more specific concept are also instances of the more general one. The alternative definition of the extension C^i_{alt} of a concept C is then defined as $\bigcup_{D \sqsubseteq C} D^i$. We will refer to a measure M based alternative extension as $Mhier$.

3.4 Calculating Mappings from Rankings

Once decided on a suitable measurement we order mappings according to their degree of relatedness. From such an ordering we can derive all sorts of mappings,

such as 1-1 or 1-n mappings.[4] In practise one also has to choose a cut-off point, *i.e.* a value of the measure, below which a mapping of two instances will be considered too unreliable.

As both the choice of cut-off and 1-n mappings is strongly application-specific, in our experiments we evaluate more generally. Instead of evaluating a particular mapping based on a particular setting we assess the quality of the ranking, *i.e.* we calculate whether a mapping suggested in a particular position in the ordering induced by the measure is correct or not.[5]

4 Experimental Setup

In our experiments we used the 5 measures described in the previous section: Jaccard, corrected Jaccard, PMI, LLR, and IGB, as well as hierarchical and non-hierarchical extensions, and two alternative thresholds (1 and 10) to deal with statistically insignificant information. Having calculated the similarity between all pairs of concepts from GTT and Brinkman we then rank these pairs of concepts based on their similarity measure in a descent order. In Section 5, we will give comparison of precision and recall of our experiments with respect to these different options.

4.1 Dataset and Types of Mappings

In our dataset, there are 243,886 books which were doubly annotated with concepts from GTT and Brinkman. In total 24,061 GTT concepts and 4,990 Brinkman concepts have been used for annotation. For each GTT and Brinkman concept, we treated the books annotated by this concept as its instances. As both ontologies are thesauri, we expect our target mapping relations, beyond the expected "equivalent to", to be the usual thesaurus semantic relations "broader than", "narrower than" and/or "related to".

4.2 Evaluation Methods

To be able to estimate the quality of a mapping, we need an evaluation procedure. For each measure we calculate four ordered lists, two taking the hierarchy into account, two not, of which one is based on a threshold of 1, and one on a threshold of 10. To get a better understanding on the difference between the measures, we calculate the overlap of the different lists. Table 1 shows the percentage of shared mappings between the ranked lists generated by all similarity measures up to the first 10,000 mappings.

[4] A 1-n mapping can be obtained as follows: for a source concept S let (S, T) be the first pair in the ordering. Then all pairs (S', T) for $S \neq S'$ are deleted from the list. 1-1 mappings can be created by deleting all (S, T') for $T \neq T'$ as well. Other cardinality choices are possible, including m-1 and m-n ("many-to-many") mappings.

[5] As an indication, we will sometimes use specific cut-offs of 100, 1,000, and 10,000 mappings, which makes the comparison of different measures easier. These numbers are relatively arbitrary, though.

Table 1. Comparison between top 10,000 mappings generated by our original measures

JC				
JC_{corr}	80%			
LLR	39%	46%		
IGB	15%	15%	9%	
MI	37%	28%	10%	10%
	JC	JC_{corr}	LLR	IGB

Table 1 shows a surprisingly big difference in the lists of mappings found using the different measures. This shows that there are indeed significantly different, and a systematic evaluation will be of crucial importance.

Due to the size and complexity of the task a complete evaluation of the correctness of the calculated mappings by domain experts is out of the question. As an alternative, we have developed an evaluation procedure consisting of three steps: **producing a Gold Standard**, calculating **average precision** and **recall approximation**. Part of this procedure is based on the simple, admittedly simplistic assumption that concepts with identical labels are equivalent.

Producing a Gold Standard. In order to evaluate the precision of the mappings generated by different measures, we first sampled the generated mappings to a reasonable size for human evaluation to produce a Gold Standard manually. For each list of mappings, we selected the top 100 mappings, every tenth mapping from the 101st to 1,000th mapping, and every 100th from 1,001st to 10,000th mappings. We filtered out all lexically equivalent mappings, since we already consider them to be valid. This produces 1,600 mappings for human evaluation.

The selected mappings were presented in random order to 3 Dutch native speakers who assigned relations "equivalent to," "broader than," "narrower than," "related to," "no link" and "do not know" to all pairs. The (online) evaluation we set out allowed evaluators to access, for the concepts involved in a mapping, both thesaurus information (*e.g.* broader concepts) and the books annotated with them.

Ordering mappings by similarity measure does not necessarily suggest an interpretation in terms of the target mapping relations "equivalent to," "broader than," "narrower than" and/or "related to." Our evaluation allows us to consider that three different types of mapping are correct: we can consider the highly ranked mappings to be

1. equivalences only (ONLYEQ),
2. equivalent, broader or narrower relations, but not related-to (NOTREL)
3. all relations except explicit non-relatedness. (ALL)

Each way of *interpreting the nature of a found mapping* will have its use in practical applications,[6] and conceptually we do not prefer any one over any

[6] Equivalence might be used, for instance, in a data translation application, while broader and narrower links can be exploited for hierarchical browsing, as in [8].

other. However, we will have to study the effect of choosing a particular semantic assumption in our experiments.

Average Precision. Since the mapping set for human judgement is only a sample of the whole generated mappings, we use the following equation to calculate the average precision up to the ith mapping:

$$P_i = \frac{N_{\text{good},i}}{N_i} \tag{4}$$

where N_i is the number of mappings which are evaluated up to ith mapping, while $N_{\text{good},i}$ is the that of mappings which are evaluated as good ones.

Recall Approximation. A preliminary experiment using string comparison over concept labels shows 2,895 exact lexical matches (2,895) between GTT and Brinkman concepts, meaning that 8.2% of GTT concepts and 55.4% of Brinkman ones have a direct equivalent form in the other thesaurus.

This is quite a significant number, especially regarding the Brinkman thesaurus. As in our case lexically equivalent concepts are considered a perfect match, we argue that the recall value on lexically equivalent concepts can be used to approximate the absolute recall. Our approximation for recall, at the ith mapping, is thus $R_i = \frac{N_{\text{lex},i}}{N_{\text{lex}}}$ where $N_{\text{lex},i}$ is the number of lexically equivalents mappings among these top i mappings, and N_{lex} is the number of all lexical equivalences between these two thesauri.

Once precision and recall are calculated, the F-measure up to ith mapping is calculated as

$$F_i = \frac{2(P_i \times R_i)}{P_i + R_i}. \tag{5}$$

4.3 Goals of the Experiments

The overall goal of our study is to improve the understanding on the role of different measures and tunings on the process of instance-based mapping. This means first and foremost answering the question whether there is a **best** combination of measure, threshold and hierarchy, which outperforms all other combinations. Furthermore, we want to better understand the influence of the choice of measure and other parameters on the mapping. All this might depend on the interpretation of the found mappings, and we will have to study the effect of the assumptions made on the nature of the relations considered to be correct.

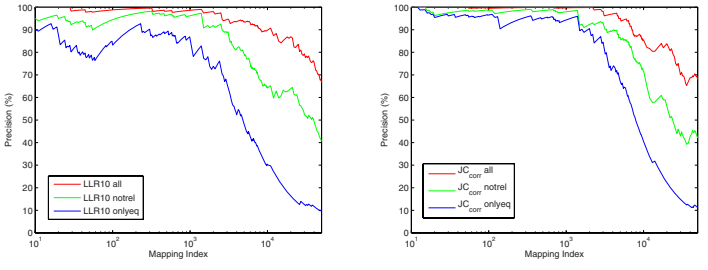
1. How does interpreting the nature of a found mapping influence results?
2. What is the influence of the choice of threshold?
3. What is the influence of using hierarchy information?
4. What is the best measure and setting for instance-based mapping?

5 Experimental Results

To answer the research questions mentioned in the previous section we performed a number of experiments in which we calculated precision, recall and f-measure.

5.1 The Influence of the Nature of a Mapping on the Results

Figure 1 shows the precision results when we use different criteria for "correctness" of mappings, *i.e.*, apart from the explicit "equivalent to" relation, whether we also count "broader than," "narrower than" or "related to" as correct mappings. As mentioned above, ONLYEQ means only those mappings which were judged "equivalent to" are counted; NOTREL counts three kinds of relation but not "related to" relation; ALL counts every relation except "no link" nor "do not know" as correct.



(a) LLR10 (b) JC_{corr}

Fig. 1. Comparing "ALL" "NOTREL" and "ONLYEQ" for LLR10 and JC_{corr}

We give the results for two measures LLR10, which is the version of LLR with threshold 10, and JC_{corr}, defining precision on the Y-axis in relation to the position in the ranking in logarithmic scale (X-axis), starting from 10 to filter the initial noise. The results for the other measures are comparable to these results. As there is a set inclusion between the sets of correct mappings, it is natural that the lines do not cross and that the top line describes the ALL-, the middle one the NOTREL- and the lower one the ONLYEQ relation.

What is more interesting is the differences between the figures. First, although LLR10 performs slightly better than JC_{corr} on the ALL counts, the precision of LLR10 is worse than that of JC_{corr} for ONLYEQ. What does this mean? It indicates that the LLR10 measure is more suitable to recognise related terms, whereas the stricter measure JC_{corr} is better at recognising proper equivalences.

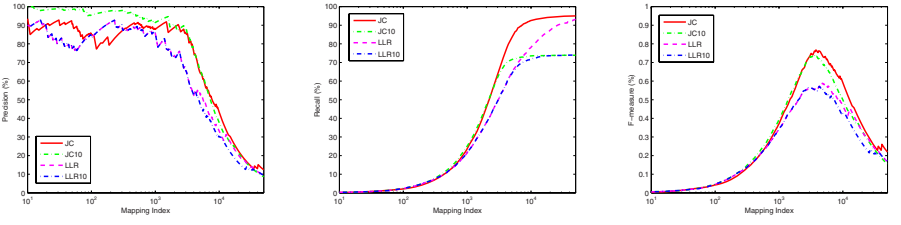
This indicates that the choice of measure has to depend on the application, and should be based on the interpretation of the nature of a found mapping. Despite the slight differences in the outcome mentioned above, we will in the following only present the results based on ONLYEQ for lack of space.

5.2 What Is the Influence of the Choice of Threshold?

An important problem in instance-based mapping is how to deal with sparse data. We have been discussing two approaches: using a threshold to exclude unreliable mappings, and statistical measures that can deal with uncertainty.

To study the effect of such a threshold, we ranked mappings according to our measures with and without a threshold. In Figure 2 we show the results for 2 measures JC and LLR, where the two dashed lines with dots are the versions with threshold, and the continuous line the Jaccard measure.

The following figures all have the same structure: the three graphs depict precision, recall and f-measure on the Y-axis versus the index of the mapping on the X-axis (which is given in a logarithmic scale).



(a) precision (b) recall (c) f-measure

Fig. 2. Comparison with respect to the threshold

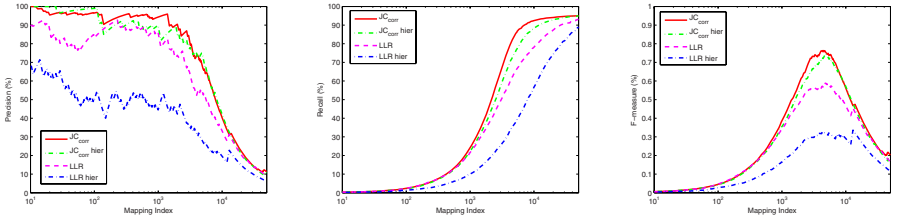
The results are in line with our expectation. As LLR has been developed for dealing with spareness, we expect the difference between the version with and without the threshold to be more similar than this is the case for Jaccard. This shows clearly in the precision, which is almost the same for LLR and LLR10. What is also expected is the significant drop in recall for both measures with threshold at around 5000 mappings. Remember that choosing a threshold simply excluded the mappings from consideration, which will also exclude many correct mappings. Also, it is interesting to notice that when considering the ALL interpretation for mappings, the gain in precision is less significant for Jaccard. This shows that using threshold rather discards related concepts, for which co-occurrence evidence, even in a small number of items, is very often reliable.

The general lesson is that including a threshold generally improves precision but that there is a price in recall to be paid.

5.3 What Is the Influence of Using Hierarchy Information?

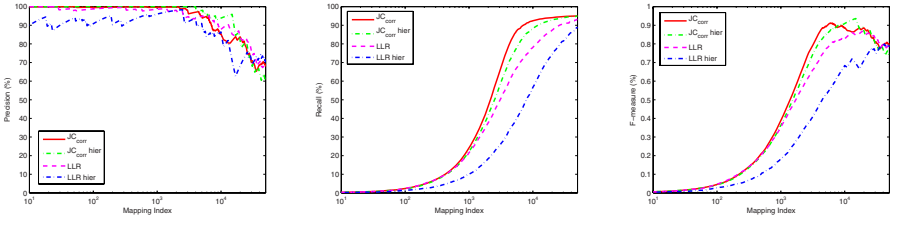
Ontology mapping is different from co-occurrence in texts in that the concepts are hierarchically organised. To find out what effect including this hierarchical information has on instance-based mapping we compared the four most promising measures with and without instances of the descendants in calculating the mappings. We also performed this experiment on different interpretations of the nature of a found mapping, and found quite diverse results.

Figure 3 shows the results of comparing JC_{corr} and LLR with their versions JC_{corr}hier and LLRhier, where we consider the ONLYEQ interpretation, *i.e.* we only consider those mappings to be correct that the evaluators have marked as equivalent. The most striking result is the gigantic drop in precision for LLRhier, as compared to JC_{corr}, for which the results are very competitive when considering



(a) precision (b) recall (c) f-measure

Fig. 3. Comparison with respect to the hierarchical information (ONLYEQ)



(a) precision (b) recall (c) f-measure

Fig. 4. Precision comparison with respect to the hierarchical information (ALL)

hierarchy. Given that we only consider equivalence statements, this shows that including instances from descendants of concepts weakens the strength of equivalent concepts in the LLR measure.

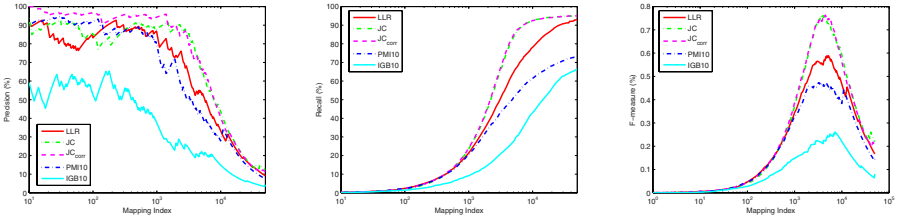
To validate this assumption, we considered the same experiments with the ALL interpretation, *i.e.* we also accept related-terms and broader/narrower-than as correct mappings.

Figure 4 shows that this assumption is correct, as the drop in precision is much smaller given the more lenient interpretation of what a mapping is. Our general conclusion regarding hierarchical information: there is no significant improvement, and in most cases even a decrease in performance. A practical reason for this can also be found in the data itself. First, GTT and Brinkman include only few hierarchical links: almost 20,000 GTT terms have no parents. Second, GTT and Brinkman are thesauri, and as such their hierarchy can be interpreted as part-whole or as domain-object links. Examples for this would be "Bible" and "Gospel according to Luke."

5.4 The Best Measure and Setting for Instance-Based Mapping

We can now finally answer the question which measure and tuning is best for instance-based mapping on our dataset. For this we considered the five measures with their ideal tuning, *i.e.* JC, JC_{corr}, LLR, PMI10 and IG10.

Figure 5 shows that the most simple measures JC and JC_{corr} have highest precision and recall at almost any mapping index, which is also reflected in the overall highest f-measures.



(a) precision (b) recall (c) f-measure

Fig. 5. Final comparison of the 5 best measures

We would like to finish the overview over our results with a general remark on the quality of our mappings. In general, apart from IGB, the results are surprisingly good, as compared to results from other ontology matching evaluations [9]. This indicates that instance-based matching is probably an easier task than structure-based or label-based mapping, but also indicates that our techniques will be suitable even in critical applications.

6 Conclusion

In this paper we presented an empirical study of instance-based matching based on a number of experiments performed on an application in the National Library of the Netherlands. We produced a Gold Standard for good mappings, and evaluated 5 different well-studied similarity measures, as well as two different ways to fine-tune them. All representations are, or course, based on Semantic Web standards.

We have to note that the complicated and very time consuming issue of evaluation was only touched marginally in the paper. Producing a gold standard is difficult and took us a long time, but the results remain sometimes controversial among domain experts. We will address the issue in more detail in future work.

The general results are very encouraging. For the first 1000 mappings the best available measure has a precision of over 90%, and at an estimated recall level of 70% we still have a precision of over 70%. Interestingly enough these results were not achieved by the complex statistical measures, but by an adapted version of the simple Jaccard measure.

The use of thresholds and hierarchical information had little influence in general, though the latter needs more study. The question here, and one that will probably apply to a number of our results, is how dependent the results are on our particular collection, and our ontologies.

For this reason we intend to conduct a similar analysis on other corpora, *e.g.* web directories or music classifications. We are confident, however, that general results are domain independent, and that instance-based mapping is a reliable and high-performing approach to ontology mapping.

Acknowledgements

STITCH is funded by NWO, the Dutch Organisation for Scientific Research. We would like to thank Henk Matthezing for his constant advice at the KB, and Claus Zinn for his valuable comments on the manuscript.

References

1. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
2. Vizine-Goetz, D.: Popular LCSH with Dewey Numbers: Subject headings for everyone. Annual Review of OCLC Research (1997)
3. Avesani, P., Giunchiglia, F., Yatskevich, M.: A large scale taxonomy mapping evaluation. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, Springer, Heidelberg (2005)
4. Isaac, A., Matthezing, H., van der Meij, L., Schlobach, S., Wang, S., Zinn, C.: The value of usage scenarios for thesaurus alignment in cultural heritage context. Under submission
5. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. Journal of data semantics 4, 146–171 (2005)
6. Doerr, M.: Semantic problems of thesaurus mapping. Journal of Digital Information 1(8) (2004)
7. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)
8. van Gendt, M., Isaac, A., van der Meij, L., Schlobach, S.: Semantic Web Techniques for Multiple Views on Heterogeneous Collections: a Case Study. In: 10th European Conference on Digital Libraries (ECDL), Alicante, Spain (2006)
9. Euzenat, J., Mochol, M., Shvaiko, P., Stuckenschmidt, H., Šváb, O., Svátek, V., van Hage, W.R., Yatskevich, M.: Results of the ontology alignment evaluation initiative 2006. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L. (eds.) ISWC 2006. LNCS, vol. 4273, Springer, Heidelberg (2006)