Cross-Lingual Data Quality for Knowledge Base Acceleration across Wikipedia Editions

Julianna Göbölös-Szabó
Hungarian Academy of Sciences
Budapest, Hungary
gszj@sztaki.hu

Natalia Prytkova, Marc Spaniol and Gerhard Weikum
Max-Planck-Institut für Informatik
Saarbrücken, Germany
{natalia|mspaniol|weikum}@mpi-inf.mpg.de

ABSTRACT

Knowledge-sharing communities like Wikipedia and knowledge bases like Freebase are expected to capture the latest facts about the real world. However, neither of these can keep pace with the rate at which events happen and new knowledge is reported in news and social media. To narrow this gap, we propose an approach to accelerate the online maintenance of knowledge bases.

Our method, called LAIKA, is based on link prediction. Wikipedia editions in different languages, Wikinews, and other news media come with extensive but noisy interlinkage at the entity level. We utilize this input for recommending, for a given Wikipedia article or knowledge-base entry, new categories, related entities, and cross-lingual interwiki links. LAIKA constructs a large graph from the available input and uses link-overlap measures and random-walk techniques to generate missing links and rank them for recommendations. Experiments with a very large graph from multilingual Wikipedia editions demonstrate the accuracy of our link predictions.

1. INTRODUCTION

1.1 Motivation

Knowledge-sharing communities like Wikipedia and knowledge bases like Freebase are thriving and keep growing at impressive rates [3]. Ideally, whenever something important happens in the real world (e.g., Mark Zuckerberg getting married), they would reflect these news and the associated entity-relationship facts already on the next day. With very prominent events, this is indeed the case. However, for events that are of regional interest only or refer to entities "in the long tail" (e.g., indie music bands), there is often a big delay before they are picked up by Wikipedia or knowledge bases. Reducing these delays by automatically generating timely and informative recommendations to Wikipedia authors and knowledge-base curators is the theme of *knowledge base acceleration* [2].

As a concrete example, consider the news announcing the ministers of the new French government under president Hollande and prime minister Ayrault. This was covered by the French Wikinews on 16-May-2012[1] but not in the English Wikinews. The French news had links to the French Wikipedia articles of all ministers. The government includes Madame Dominique Bertinotti as minister for family affairs reporting to the prime minister. She has an extensive French Wikipedia page, but is not covered by any other Wikipedia edition. Another example is Nicole Briq, the minister for environment and sustainable energy. She has an English Wikipedia page, but it merely consists of a single short paragraph, a so-called "stub page" with the comment that it needs to be expanded. Interestingly, there is no German Wikipedia page about Madame Briq, although Germany is a neighboring country of France. All this is as of 23-May-2012, a full week after the original news in France, and despite the fact that English and German media certainly reported about the new government as well.

In this scenario, the idea of knowledge base acceleration would be to give recommendations to the non-French communities about interwiki links that should be established between pages of different language editions, categories into which new or expanded articles should be placed, and also interwiki links between categories. For example, the English page about Nicole Briq lists only 6 categories, whereas the French article has almost twice as many categories. However, the interwiki linkage between categories is sparse and often noisy. For example, Ayrault (the prime minister) is in the French category "Maire de Nantes", but this category is not linked to the existing English category "Mayors of Nantes". Such recommendations for additional articles, categories, and links should be generated in an automated manner by analyzing several Wikipedia editions across languages and by considering online news that mention Wikipedia entities in at least one language (ideally in the form of hyperlinks or linked-data formats like RDFa [1] statements).

1.2 Contribution

This paper addresses the outlined problem of generating recommendations for knowledge base acceleration. Recommending missing links, like interwiki links or membership in categories, resembles the well studied link prediction problem [10, 12]. However, prior work on this topic has focused on friendship relations in social networks or on product recom-

[1] http://fr.wikinews.org/wiki/France_:_annonce_de_la_composition_du_gouvernement_Ayrault?dpl_id=43909

mendations, whereas our setting needs to reconcile highly heterogeneous nodes with rich contents in different languages as well as noisy nodes like news articles. The mixed quality of interwiki links across Wikipedia editions has been discussed in prior work as well (e.g., [13, 6]). However, that work only considered spurious links between non-matching articles of different languages, but none of the prior methods considered generating entirely new interwiki links between contents-rich articles in one edition and sparse articles or stub pages in another edition.

Our approach is based on a graph model for reconciling the different kinds of nodes and links that we obtain from multi-lingual Wikipedia editions. We have so far concentrated on three languages: French, German, and Hungarian, considering two large and one small knowledge-sharing community. For recommending links in this heterogeneous graph, we use link-overlap measures such as weighted Jaccard and random-walk techniques such as SimRank [9]. We generalize the notion of SimRank to work with our knowledge graph model, introducing a weighted-edges extension of SimRank.

The main contributions of the paper are as follows:

- We introduce a knowledge graph model over multi-lingual editions of Wikipedia (and potentially other sources), and we define three types of link recommendation problems.
- We generalize the notion of SimRank to address all three recommendation types and develop a suite of efficient algorithms for predicting and ranking missing links in the knowledge graph;
- We report on experimental studies with Wikipedia editions in three different languages.

2. GRAPH MODEL

Consider Wikipedia articles from different language editions, and possibly even different versions in the Wikipedia history. Each article or article version is a node in a graph that we build across multiple editions. The edges in this graph comprise both hyperlinks within one Wikipedia edition as well as interwiki links between different editions.

Each node in the graph has

- a *name*: the unique name of the article as of today;
- its *language*: the Wikipedia edition to which the article belongs;
- an *identifier*: composed of name and language;
- a set of *outgoing links*: both within and across editions, including links to articles as well as Wikipedia categories;
- a set of *incoming links*: from articles and categories or interwiki links.

Wikipedia is coupled with large entity-relationship-structured knowledge bases like dbpedia.org or yago-knowledge.org, which have been built from Wikipedia and further sources and captured their Wikipedia provenance. This way, Wikipedia articles in English and other languages can be associated with entities in DBpedia [4] or Yago [17]. This in turn allows us to annotate nodes with additional properties. Our setting considers article and category nodes.

Once we have different sorts of nodes, we can distinguish different kinds of edges in our graph model:

- links between articles of the same Wikipedia edition;
- links between articles and categories to which they are assigned;
- links between categories and super-categories or subcategories;
- interwiki links that go across two Wikipedia editions, connecting either article pairs or category pairs (but never any mixed pair like article-category).

This graph construction gives us an expressive model for mining the data quality across multiple Wikipedia editions. The semantic enhancements via Yago and the temporal support can be used to focus on specific slices of Wikipedia, for example, on the cross-lingual quality for the political domain (types politicians, parties, governments, etc.) as of a time period around the last election in France. In this paper, we will not consider such advanced analyses, though, leaving them to future work. Instead we will consider only the current snapshot of Wikipedia-derived nodes, ignoring the Yago-based annotations.

Figure 1 illustrates our graph model. In the rest of the paper, we will use the following notation. Nodes of sort article or category, in language edition l are denoted as $a_i^{(l)}$ or $c_i^{(l)}$, respectively. We simply write c_i and a_i if the language is clear from the context, and we write v_i if the node (vertex) sort is irrelevant Analogously, we speak of edge sorts a-a, a-c, $a^{(l)}$-$a^{(\tilde{l})}$ etc. $Art^{(l)}(c_i^{(l)})$ denotes a set of articles of category $c_i^{(l)}$ w.r.t. language l. $Cat^{(l)}(a_i^{(l)})$ returns a set of categories, article $a_i^{(l)}$ belongs to. $I^{(\tilde{l})}(x_i^{(l)})$ gives the equivalent node in language $\tilde{l}$ for an article or category $x_i^{(l)}$ in language l.

3. LINK RECOMMENDATION

3.1 Types of Recommendations

We address three different link recommendation types:

1: **Type-1 recommendation:** new interwiki link for category $c_i^{(l)}$. Recommend an interwiki link between categories in different languages.

2: **Type-2 recommendation:** new category $c^{(l)}$ for article $a_i^{(l)}$. Recommend links between articles and categories in the same language.

3: **Type-3 recommendation:** related articles (entities) $a^{(l)}$ for article $a_i^{(l)}$. Recommend links between an article and related articles in the same language.

We focused on these three types of ranked link predictions, because we consider them being the bottlenecks in cross-lingual data quality and knowledge base acceleration. We actually disregard the most obvious prediction type: recommending interwiki links between articles, because Wikipedia editions already exhibit very high coverage and accuracy in this regard. For categories, however, the interwiki linkage is much sparser and noisier. Recommendation types 2 and 3 aim to support a Wikipedia author by making suggestions for categories into which a new article or the extension of a stub page should be placed and for related entities that should be mentioned in the new article. Although types 2 and 3 have both input (an article) and output (categories or related articles) in the same language, the point of these recommendations is to utilize the linkage with and contents of other languages – just that there is no immediate linkage available, so detours must be considered.

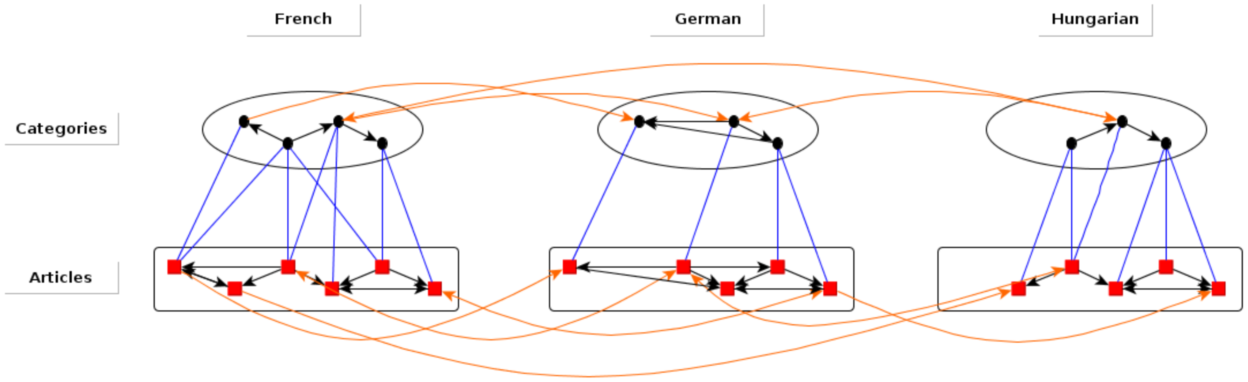


Figure 1: Illustration of the Cross-Lingual Graph Model

3.2 Algorithms

For each of the three types, the algorithmic framework is similar. For a given start node $v_i^{(l)}$, we compile a set of *potential targets*, $Cand(v_i^{(l)})$, by traversing the cross-lingual graph with a bounded number of hops. Then we use set-overlap or random-walk techniques to score and rank the candidates. Note, that for all scoring techniques the candidate set for a particular type of prediction remains the same.

Algorithms 1 and 2 illustrate individual steps of type-1 and type-2 predictions. The algorithm for type-3 recommendation is analogous to a type-2 prediction, except that in step 4 we consider *a*-*a* links instead of *a*-*c* edges. For ranking of potential recommendations we studied various methods:

- The *overlap* is specified between the neighbor set of a candidate $x \in Cand(v_i^{(l)})$ and the original node $v_i^{(l)}$. For type-1 where we start with a category $c_i^{(l)}$ and consider candidates $x = c_j^{(\tilde{l})}$. The neighbor sets are simply the articles in $c_i^{(l)}$ and $c_j^{(\tilde{l})}$, restricted to those articles that have bidirectional interwiki links between l and $\tilde{l}$. For type-2 we start with an article $a_i^{(l)}$ and consider candidates $x = c_j^{(l)}$. These are the neighbor sets of language-l articles that can be reached from a_i and c_j, respectively, a) by a single step to neighboring articles, or b) by a two-step walk from article to category and its member articles, or c) by a three-step walk from an l node to the interwiki neighbor in $\tilde{l}$, looking up member articles or categories in $\tilde{l}$, going back to their counterparts in l via bidirectional interwiki links (not counting as a hop in the walk), and looking up member articles or categories in l. For type-3 we start with an article $a_i^{(l)}$ and consider related articles as candidates $x = a_j^{(l)}$. Here, the neighbor set is constructed by looking up the parallel-language counterpart of $a_i^{(l)}$, say $b_i^{(\tilde{l})}$, collecting all article neighbors of $a_i^{(l)}$ and $b_i^{(\tilde{l})}$ that are reached by outgoing links, and mapping the $\tilde{l}$ articles back to l by the available interwiki links.
- The *similarity* of a candidate x and the originally given node $v_i^{(l)}$ is defined under the extended notions of the *SimRank* measure [9]. See Subsection 3.4 for further discussion. This method works uniformly for all recommendation types, as it can applied to any pair of nodes in our graph model.
- An overlap-based *voting* scheme for the type-1 recommendation problem (see below).
- A *novelty-oriented variant of SimRank* tailored for the type-2 recommendation problem (see below).

Algorithm 1 Type-1 recommendation

```
1: procedure RECOMMEND(category $c_i^{(l)}$)
2:     $Cand(c_i^{(l)}) = \emptyset$
3:     for all $a_i^{(l)}$ in $Art^{(l)}(c_i^{(l)})$ do
4:         $\tilde{a_i}^{(\tilde{l})} = I^{(\tilde{l})}(a_i^{(l)})$
5:         $Cand(c_i^{(l)}) = Cand(c_i^{(l)}) \cup Cat^{(\tilde{l})}(\tilde{a_i}^{(l)})$
6:     end for
7:     rank $Cand(c_i^{(l)})$ wrt similarity to $c_i^{(l)}$
8:     return $Cand(c_i^{(l)})$
9: end procedure
```

Algorithm 2 Type-2 recommendation

```
1: procedure RECOMMEND(article $a_i^{(l)}$)
2:     $\tilde{a_i}^{(\tilde{l})} = I^{(\tilde{l})}(a_i^{(l)})$
3:     $Cand(a_i^{(l)}) = \emptyset$
4:     for all $c_i^{(\tilde{l})}$ in $Cat^{(\tilde{l})}(\tilde{a_i}^{(\tilde{l})})$ do
5:         if $\nexists I^{(l)}(c_i^{(\tilde{l})})$ then
6:             continue
7:         end if
8:         if $I^{(l)}(c_i^{(\tilde{l})}) \in Cat^{(l)}(a_i^{(l)})$ then
9:             continue
10:        end if
11:        $Cand(a_i^{(l)}) = Cand(a_i^{(l)}) \cup \{I^{(l)}(c_i^{(\tilde{l})})\}$
12:    end for
13:    rank $Cand(a_i^{(l)})$ wrt similarity to the set $Cat^{(l)}(a_i^{(l)})$
14:    return $Cand(a_i^{(l)})$
15: end procedure
```

3.3 Overlap-based Methods

Jaccard-based methods: All overlap measures are implemented by *Jaccard coefficients* over the respective neighbor sets. In addition to standard Jaccard

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

we also use *weighted Jaccard* [8] where weights of categories (as elements in the neighbors sets) are proportional to category sizes and weights of articles (as elements of neighbors sets) are proportional to the numbers of categories to which they belong:

$$wJ(A,B) = \frac{\sum_{e \in A \cap B} min(weight_A(e), weight_B(e))}{\sum_{e \in A \cup B} max(weight_A(e), weight_B(e))}$$

Voting method: In addition, we devise a simpler method of overlap-based *voting*. For the recommendation of interwiki links between categories in languages l (starting point) and $\tilde{l}$ (target), a simple heuristics is to directly compare the article sets $Art^{(l)}(c_i^{(l)})$ and the various candidates x in $\tilde{l}$: $Art^{(l)}(x^{(\tilde{l})})$ where $\tilde{l}$ members are restricted to those who have bidirectional interwiki links with l articles and can thus be trivially mapped backed to l. The voting method prefers the category x with the largest number of articles shared between $Art^{(l)}(c_i^{(l)})$ and $Art^{(l)}(x^{(\tilde{l})})$.

3.4 SimRank-based Methods

Standard SimRank [9] is a widely used measure of structural context similarity. A multilingual Wikipedia graph has rich interlinkage, which we utilize while computing SimRank value. The SimRank measure between nodes v and x is defined in the following (recursive) way, with a decay factor $\gamma (0 < \gamma < 1)$ and $In(v)$ denoting the set of inlink neighbors of node v:

$$SR(v,x) = \begin{cases} \frac{\gamma}{|In(v)|\ |In(x)|} \sum_{v' \in In(v)} \sum_{x' \in In(x)} SR(v', x') & \text{if } v \neq x \\ 1, \text{otherwise} \end{cases}$$

It has been shown in [7] that $SR(v,x)$ is equivalent to the expected length (i.e., number of hops) of the first meeting of two coupled random walks, one starting from v and one starting from x. This gives rise to a highly efficient computation [7]: a) compute (standard) random walks of bounded lengths with each node as a starting point; b) organize the reached nodes and their meeting distances in a *fingerprint tree (FPT)* [7], c) repeat the random walks with different random choices, creating m i.i.d. (independently identically distributed) FPTs. All this is precomputed in time $O(l \cdot m \cdot N)$, where l is the length of the random walks, m is the number of i.i.d. repetitions, and N is the number of nodes in the graph. Later, when we want to know the SimRank measure for two nodes, we only have to find the nodes in each of the m FPTs and look up the distance (number of hops) until the walks meet. The distances are then averaged over all m FPTs, thus approximating (and converging to) the expected meeting distance. This online procedure has time complexity $O(l \cdot m)$.

Extended SimRank: We have extended the notion of SimRank by allowing weights for all edges as follows:

$$wSR(v,x) = \gamma \sum_{v' \in In(v)} \sum_{x' \in In(x)} w(v' \to v) w(x' \to x) SR(v', x')$$

The weighting allows to reduce bias towards highly populated categories or articles with a large number of links. We have shown that this extension is still equivalent to the expected meeting distance for the corresponding weighted (i.e., non-uniform) coupled random walks. The theorem and its proof are omitted for space limitation.

Another deviation from standard SimRank is that we allow also conceptual self-loops, introducing a bias towards reaching local-neighborhood nodes (i.e., penalizing long-distance walks). For this purpose, we introduce parameter ϵ for a random-walk hop staying in the same node. Different edge weights are used for the different link types: a) *a-a* edges in the same language, b) *a-c* edges in the same language, c) *a-a* edges or *c-c* edges across languages. We introduce α, β, ξ for the probabilities of following each of the three edge types stated above. It would be easy to further refine the notion of edge weights to accomodate the size of categories or the length of articles to bias the choice among link destinations. Our experiments use only edge-type-specific weights, though.

Novelty ranking: For the recommendation of language-l categories for a given article $a_i^{(l)}$, we would ideally like to rank categories that are not obvious and most novel relative to the known categories $Cat^{(l)}(a_i^{(l)}) = \{c_1, c_2, \dots\}$. For example, if we already know that Jean-Marc Ayrault is in the category "Prime Ministers of France", and we obtain, via other languages, candidate categories "Members of the French Socialist Party" or "teachers", we prefer the latter because the former has high overlap with "Prime Ministers of France" and other already known categories. Intuitively we want to achieve a high information gain, or novelty. We implemented this idea by extending SimRank into a notion SR^* of $n+1$ coupled random walks. We consider the random walks at starting nodes $c_1, c_2, \dots, c_n$, the known categories, and a recommendation candidate x. We compute the expected length until all $n+1$ walks meet, using the fingerprint trees, and then define

$$Novelty(x) = 1 - SR^*(c_1, c_2, \dots, c_n, x)$$

We have shown that SR^* is equivalent to an n-way generalization of the recursive SimRank definition. The computation determines the maximum distance over all $n-1$ meeting points for each fingerprint tree, and then averages over all trees. Details are omitted for space limitation.

4. THE LAIKA SYSTEM

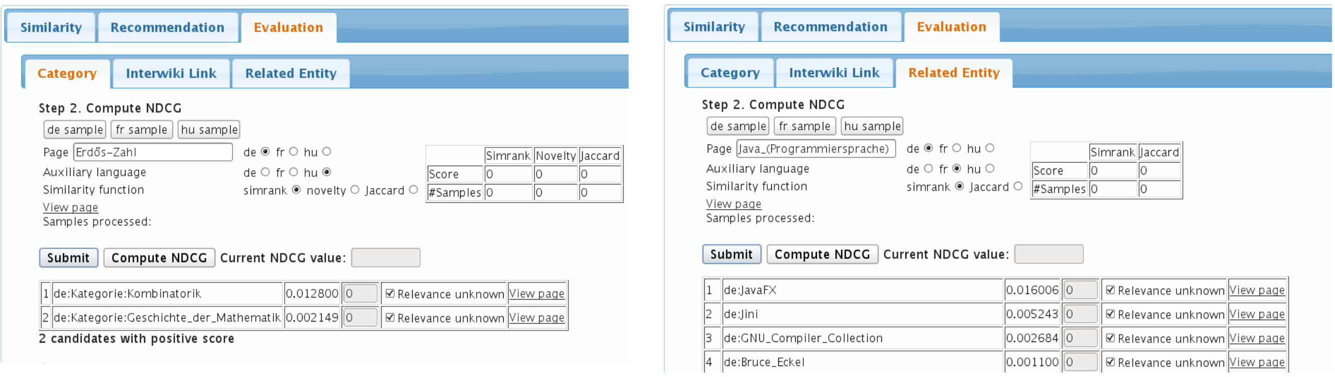
All components of the LAIKA system (short for Link AntIcipation for Knowledge-base Acceleration) are implemented in C++. LAIKA has a client and a server part. The client provides a Web interface, supporting three modes: a) similarity computations between given articles and/or categories using a variety of our methods; b) recommendations for the three link-prediction types; c) an evaluation mode where users can assess the quality of results. Figure 2 shows the user interface in this mode. The server part is responsible for the computations on the multilingual Wikipedia graph.

5. EXPERIMENTS

5.1 Setup

We have downloaded the complete Wikipedia editions for German, French, and Hungarian, as of March 2012. This choice was made to capture two of the larger Wikipedia editions, German and French which have similar sizes, and one smaller edition, Hungarian. We considered a pair of pages in two parallel editions to be equivalent if the pages were connected via interwiki links in both directions. Tables 1 and 2 summarize the resulting datasets and their cross-linkage.

For the SimRank-based methods we computed 400 fingerprints (iid random walks) of maximum length 100. In total, for all 5.5 Million nodes in our graph, this precomputation took ca. 3 hours on a simple Linux server.



Step 2. Compute NDCG

Page Erdős–Zahl de ● fr ○ hu ○
Auxiliary language de ○ fr ○ hu ●
Similarity function simrank ● novelty ○ Jaccard ○
Samples processed:

	Simrank	Novelty	Jaccard
Score	0	0	0
#Samples	0	0	0

Current NDCG value:

1	de:Kategorie:Kombinatorik	0.012800	0	☑ Relevance unknown	View page
2	de:Kategorie:Geschichte_der_Mathematik	0.002149	0	☑ Relevance unknown	View page

2 candidates with positive score

Step 2. Compute NDCG

Page Java_(Programmiersprache) de ● fr ○ hu ○
Auxiliary language de ○ fr ● hu ○
Similarity function simrank ● Jaccard ○
Samples processed:

	Simrank	Jaccard
Score	0	0
#Samples	0	0

Current NDCG value:

1	de:JavaFX	0.016006	0	☑ Relevance unknown	View page
2	de:Jini	0.005243	0	☑ Relevance unknown	View page
3	de:GNU_Compiler_Collection	0.002684	0	☑ Relevance unknown	View page
4	de:Bruce_Eckel	0.001100	0	☑ Relevance unknown	View page

Figure 2: LAIKA Web interface for evaluation. Type-1 on the left and type-3 on the right.

Language	# articles	# categories	# links
De	2 338 795	139 844	45 531 135
Fr	2 408 097	199 708	42 022 704
Hu	339 041	34 653	6 273 337

Table 1: Sizes of Wikipedia editions in March 2012.

Equivalent articles		Equivalent categories	
De-Fr	482 196	De-Fr	22 175
De-Hu	108 949	De-Hu	4 840
Fr-Hu	119 559	Fr-Hu	5 387

Table 2: Interwiki links in March 2012.

Experiments with Ground Truth. For each of the three link-prediction types, we generated a test case with well-defined ground truth as follows. We randomly removed 10% of the interwiki links (between two languages), the article-category links (in the target language l), and article-article links (in the target language l), respectively. Then we predicted and ranked missing links, and compare the ranked results of the different recommendation methods against the originally existing links. This way we could automatically compute standard measures for the output quality:

- **MRR (mean reciprocal rank):** the reciprocal of the highest rank at which a correct result appears (a standard measure for recommender problems);
- **Recall:** the fraction of ground-truth links recommended by a method;
- **Precision:** the fraction of correct links among the recommended ones;
- **Precision@10:** the precision for the top-10 ranks only;
- **NDCG (normalized discounted cumulative gain):** the accumulated precision over all ranks, with ranks weighted in a geometrically decreasing manner (a standard measure for rankings in IR).

Manual assessment. The type-2 and type-3 recommenders can return relevant categories or articles that are not among the 10% removed links for the ground-truth construction. Thus, we also performed a manual assessment, using an evaluation tool (see Figure 2) and human judges. We generated random samples of German pages and used the French Wikipedia as the parallel corpus for recommending *a*-*c* and *a*-*a* links in German.

	MRR	NDCG	Recall	Precision	Prec10
Weighted Jaccard	0.539	0.764	0.630	0.214	0.227
Extended SimRank	0.518	0.645	0.630	0.214	0.219
Voting	0.712	0.850	0.630	0.214	0.230

Table 3: Results for type-1: interwiki links.

5.2 Results

Type-1 recommendations: interwiki links. Table 3 shows the results for three different recommendation methods. Note that in type-1 recommendation, there is only one correct result (the equivalent category in the parallel language). Therefore, there is no point in evaluating precision; instead MRR and NDCG (with weighted ranks) are the main measures of interest. All results in the table are averaged over all instances (among the 10% removed links) for all 6 language pairs: 13,000 links for type-1, 914,000 for type-2, 8.5 million for type-3.

We observed that all methods performed extremely well, with the Voting method excelling. An MRR value above 0.5 means that, on average, the correct result was found on rank 1 or 2 – in other words, nearly perfect predictions. The recall value of 0.63 tells us that in more than half of the instances, we found the correct missing link (not necessarily always in the top ranks, though). An example of a type-1 recommendation is: for the German category *Seltsame Materie* (engl.: strange matter), with only 9 pages, we recommended the Hungarian category *Csillagászati alapfogalmak* (engl. basic astronomical concepts). Note that these are small categories in the long tail. For example, the Hungarian Wikipedia does not contain a category on strange matter at all. So the recommendations are not obvious, and the high accuracy of our methods is remarkable.

Type-2 recommendations: new categories. In this case, the recommenders can produce multiple correct outputs. Thus, precision and precision@10 for the scored and ranked categories is interesting. MRR refers to the rank of the highest-ranked correct result; NDCG reflects all correctly predicted positions in a ranking. Table 4 shows the results, comparing the weighted Jaccard, the extended SimRank, and the Novelty methods. Again, the MRR and NDCG values are extremely good; so we recommend correct categories at ranks 1 or 2 in most cases. For this task, our Novelty method outperformed the other methods by a significant margin.

The recall is the same for all methods, as they worked on the same candidate sets, solely ranking them differently. The

	MRR	NDCG	Recall	Precision	Prec10
Weighted Jaccard	0.734	0.857	0.367	0.291	0.291
Extended SimRank	0.757	0.883	0.367	0.291	0.291
Novelty	0.762	0.910	0.367	0.291	0.291

Table 4: Results for type-2: categories.

	MRR	NDCG	Recall	Precision	Prec10
Weighted Jaccard	0.787	0.539	0.165	0.062	0.068
Extended SimRank	0.781	0.518	0.165	0.062	0.065

Table 5: Results for type-3: related articles.

recall number of ca. 36% tells us that the recommenders still miss out on many correct results. This is due the fact that many candidates were assigned a score of zero, when overlap measures were zero or the coupled random walks did not result in meetings at all. For the Novelty method, this effect also led to many ties in the scoring (of seemingly perfect score 1), which were then arbitrarily ordered. All SimRank-based methods could potentially overcome this limitation in recall, by increasing the number of precomputed FPTs (iid walks).

An example of a type-2 recommendation is: for the article *Kosmische Strahlung* (engl.: cosmic ray) with the help of the French edition, our methods suggested the categories: *Astrophysik* (astrophysics), *Elektromagnetisches Spektrum* (electromagnetic spectrum), *Teilchenphysik* (particle physics).

As for the manual assessment recommended categories, we observed an average precision of 86% for the extended SimRank method and 84% for the Novelty method. These numbers refer to all candidate categories ranked by our methods with scores greater than zero.

Type-3 recommendations: related articles. For type-3 recommendation, the situation is similar to type-2 except that the numbers of candidates, related articles in this case, is usually much higher than for type-2 category recommendation. The results are shown in Table 5. Again, the MRR and NDCG numbers demonstrate the high quality of our methods, with weighted Jaccard slightly outperforming the extended SimRank. The precision numbers are fairly low: our methods picked up many remotely related articles such as year or country pages for people as targets. This illustrates the potential of connecting our graph model with a semantic type system like the Yago classes; we could then easily filter out recommended articles that do not fit a given type profile (e.g., filter everything out but people and organizations). The recall numbers are also smaller than for the type-2 case. Here, the much larger candidate sets aggravated the problem of zero overlap or non-meeting random walks.

An example of a type-3 recommendation is: for the German article *Kosmische Strahlung* (cosmic ray), related articles about people include *Pierre Auger*, *Arthur Holly Compton* and *Charles Thomson Rees Wilson*. All of them are famous physicists in the field of nuclear and cosmic ray physics. Other related articles are *Teilchenphysik* (particle physics), *Elementarteilchen* (elementary particles), *Strahlung* (radiation) and *Partikel* (particle).

Parallel languages. We also investigate the influence on the choice of language pairs for the resulting output quality. Table 6 shows the NDCG values for all recommendation types for each of the six language pairs. In the table, the target language (which the input article and the output category belong to) is in boldface (on the left), and the parallel language for generating recommendations is in normal font. Results are based on the Extended SimRank method.

	De - Fr	**De** - Hu	**Fr** - De	**Fr** - Hu	**Hu** - De	**Hu** - Fr
Type 1	0.60	0.69	0.68	0.76	0.60	0.56
Type 2	0.94	0.94	0.80	0.92	0.88	0.94
Type 3	0.56	0.46	0.51	0.47	0.45	0.48

Table 6: NDCG results for parallel languages.

6. RELATED WORK

Link Prediction in Wikipedia. [5] finds semantic relations between Wikipedia categories based on interlinkage of pages belonging to categories – all within a single Wikipedia edition. [19] addresses the problem of automatically generating links between Wikipedia articles, using NLP and learning techniques, but does not consider cross-lingual issues at all. [16] addresses the problem of missing links by tensor factorization. While performing well on highly connected graph nodes, the approach disregards nodes with low connectivity. In contrast our work is specifically geared for long-tail articles and small categories with sparse linkage.

Multilingual Wikipedia. [6] uses LP and other optimization methods for cleaning the interwiki graph across many languages. The focus is on removing spurious links and identifying sound equivalence classes of articles in parallel languages. In contrast, we address the recommendation of so far non-existing links, especially in the long tail of articles and categories. [15] considers a pair of Wikipedia editions to detect missing cross-language links between articles. The solution involves SVM classification, using a variety of link and contents features. In contrast, our method is able to predict all kinds of missing links, most notably, interwiki links for small categories and categories for long-tail articles. Also, our techniques are very efficient, compared to the time-consuming methods of [15]. [14] automatically matches infobox schemas across multiple languages in Wikipedia. This data-integration task is very different from our mission of link prediction for knowledge base acceleration. [18] pursues another data-integration problem by connecting articles from the Chinese online community Baidu Bake to the English Wikipedia. It uses a factor-graph learning method over rich contents features.

Large-scale Similarity Computation. [11] proposes a method for estimating the number of iterations that SimRank should use given a desired accuracy. The method is computationally expensive even for small graphs and not viable on a Wikipedia-scale multi-million-node graph. [7] developed the method of fingerprint trees that we build on in our work. This prior work considered standard SimRank only, whereas we devise extensions of SimRank.

7. CONCLUSIONS AND FUTURE WORK

This paper presented our recent results on understanding the data quality in cross-lingual Wikipedia editions, and on automatically generating recommendations to Wikipedia authors of different languages. Experiments with three types of link recommendation problems indicate that there is great potential for helping authors in dealing with long-tail entities and events. This work is part of an ongoing project on the overriding objective of knowledge base acceleration. Next steps include accomodating the Wikipedia history, considering external references in news and social media, and further extending our suite of recommendation methods and tools.

Acknowledgements

This work has been supported by the Hungarian Agency for National Development through the research grant TÁMOP (contract no. 4.2.2.B-10/1–2010-0009) and by the 7^{th} Framework IST programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) under contract no. 258105.

8. REFERENCES

[1] Rdfa Primer: bridging the human and data webs. W3C working group note, 14 October 2008, `http://www.w3.org/TR/xhtml-rdfa-primer/`

[2] TREC Knowledge Base Acceleration. `http://trec-kba.org`

[3] Wikistats: Wikimedia Statistics. `http://stats.wikimedia.org`

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z.G. Ives. DBpedia: A Nucleus for a Web of Open Data. In *ISWC/ASWC*, pages 722–735, 2007. Online access at `http://dbpedia.org`.

[5] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantics relationships between wikipedia categories. In *SemWiki*: First Workshop on Semantic Wikis, Budva, Montenegro, 2006.

[6] G. de Melo and G. Weikum. Untangling the cross-lingual link structure of wikipedia. In *ACL*, pages 844–853, 2010.

[7] D. Fogaras and B. Rácz. Scaling link-based similarity search. In *WWW*, pages 641–650, 2005.

[8] P. Gamallo, C. Gasperin, A. Agustini, and G.P. Lopes. Using Syntactic Contexts for Measuring Word Similarity. 4th International Conference on Text, Speech and Dialogue (TSD), Zelezna Ruda, Czech Republic, 2001.

[9] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD*, pages 538–543, 2002.

[10] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. In *J. Am. Soc. Inf. Sci. Technol.* 58(7), pages 1019–1031, 2007.

[11] D. Lizorkin, P. Velikhov, M. Grinev, and D. Turdakov. Accuracy estimate and optimization techniques for simrank computation. In *PVLDB* 1(1), pages 422–433, 2008.

[12] G. Namata and L. Getoor. Link prediction. In *Encyclopedia of Machine Learning*, pages 609–612. 2010.

[13] F. Naumann. Dr. crowdsource or how I learned to stop worrying and love web data, Keynote at 2nd International Workshop on Business Intelligence and the Web (in conjunction with EDBT 2011), Uppsala, Sweden, 2011, `http://www.hpi.uni-potsdam.de/fileadmin/hpi/FG_Naumann/publications/2011/BEWEB_2011_Invited_Talk.pdf`

[14] T. H. Nguyen, V. Moreira, H. Nguyen, H. Nguyen, and J. Freire. Multilingual schema matching for wikipedia infoboxes. In *PVLDB* 5(2), pages 133-144, 2011.

[15] P. Sorg and P. Cimiano. Enriching the crosslingual link structure of wikipedia - a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, 2008.

[16] S. Spiegel, J. Kunegis, J. Clausen, and S. Albayrak. Link prediction on evolving data using tensor factorization. In *Proceeding of ACM International Workshop on Behavior Informatics (in conjunction with PAKDD 2011)*, pages 100–110, New York, USA, 2011.

[17] F. Suchanek, G. Kasneci, G. Weikum. Yago: a core of semantic knowledge. In *WWW*, pages 697–706, 2007. Online access at `http://yago-knowledge.org`.

[18] Z. Wang, J. Li, Z. Wang, and J. Tang. Cross-lingual knowledge linking across wiki knowledge bases. In *WWW*, pages 459–468, 2012.

[19] F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *CIKM*, pages 41–50, 2007.