Alignment of Multi-Cultural Knowledge Repositories

A dissertation submitted towards the degree
Doctor of Engineering
of the Faculty of Mathematics and Computer Science of
Saarland University

by

Natalia Boldyrev
(née Prytkova)

Saarbrücken
June 2017

Dean:	Prof. Dr. Frank-Olaf Schreyer
Colloquim:	16 October 2017

Examination Board

Supervisor and First Reviewer:	Prof. Dr. Gerhard Weikum
Second Reviewer:	Prof. Dr. Klaus Berberich
Third Reviewer:	Prof. Dr. Marc Spaniol
Chairman:	Prof. Dr. Dietrich Klakow
Research Assistant:	Dr. Daria Stepanova

Acknowledgements

First of all I express my special gratitude to my supervisor Prof. Dr. Gerhard Weikum for giving me the opportunity to pursue my doctoral studies, for his invaluable contribution and many discussions we had throughout my PhD. His guidance and thorough feedback made this dissertation possible.

I sincerely thank Prof. Dr. Marc Spaniol. He was always ready to give me a helping hand, and I could always count on his support and insightful advises. I also thank him for his comments on this manuscript.

I am very grateful to Dr. Klaus Berberich for accepting my request to review the thesis and his feedback on the earlier draft of this dissertation.

I would like to thank my co-authors Julianna Göbölös-Szabó and Dr. Jannik Strötgen with whom I collaborated on parts of this work.

I deeply appreciate the help of Dr. Daria Stepanova, Dr. Rishiraj Saha Roy, Dr. Luciano del Corro, Dr. Andrew Yates, and Alkhazur Manakov with reviewing earlier drafts of this dissertation.

I acknowledge with gratitude the excellent working environment, which would not be possible without my colleagues at the Databases and Information Systems group. My sincere thanks also go to Petra Schaaf, Daniela Alessi, Stefanie Jörg and Jennifer Gerling for their assistance in many organizational issues. I sincerely thank the International Max Planck Research School for Computer Science for the financial support.

Finally, I extend my deepest thank to my family for sharing with me all the cheerful and stressful periods of my PhD. To my mother Irina for her love, care and confidence in me, to my husband Artem for his endless patience, spiritual support and encouragement, and to my son David for giving a new dimension and a new value to my life.

Abstract

The ability to interconnect multiple knowledge repositories within a single framework is a key asset for various use cases such as document retrieval and question answering. However, independently created repositories are inherently heterogeneous, reflecting their diverse origins. Thus, there is a need to align concepts and entities across knowledge repositories. A limitation of prior work is the assumption of high affinity between the repositories at hand, in terms of structure and terminology.

The goal of this dissertation is to develop methods for constructing and curating alignments between multi-cultural knowledge repositories. The first contribution is a system, ACROSS, for reducing the terminological gap between repositories. The second contribution is two alignment methods, LILIANA and SESAME, that cope with structural diversity. The third contribution, LAIKA, is an approach to compute alignments between dynamic repositories.

Experiments with a suite of Web-scale knowledge repositories show high quality alignments. In addition, the application benefits of LILIANA and SESAME are demonstrated by use cases in search and exploration.

Kurzfassung

Die Fähigkeit mehrere Wissensquellen in einer Anwendung miteinander zu verbinden ist ein wichtiger Bestandteil für verschiedene Anwendungsszenarien wie z.B. dem Auffinden von Dokumenten und der Beantwortung von Fragen. Unabhängig erstellte Datenquellen sind allerdings von Natur aus heterogen, was ihre unterschiedlichen Herkünfte widerspiegelt. Somit besteht ein Bedarf darin, die Konzepte und Entitäten zwischen den Wissensquellen anzugleichen. Frühere Arbeiten sind jedoch auf Datenquellen limitiert, die eine hohe Ähnlichkeit im Sinne von Struktur und Terminologie aufweisen.

Das Ziel dieser Dissertation ist, Methoden für Aufbau und Pflege zum Angleich zwischen multikulturellen Wissensquellen zu entwickeln. Der erste Beitrag ist ein System names ACROSS, das auf die Reduzierung der terminologischen Kluft zwischen den Datenquellen abzielt. Der zweite Beitrag sind die Systeme LILIANA und SESAME, welche zum Angleich eben dieser Datenquellen unter Berücksichtigung deren struktureller Unterschiede dienen. Der dritte Beitrag ist ein Verfahren names LAIKA, das den Angleich dynamischer Quellen unterstützt.

Unsere Experimente mit einer Reihe von Wissensquellen in Größenordnung des Web zeigen eine hohe Qualität unserer Verfahren. Zudem werden die Vorteile in der Verwendung von LILIANA und SESAME in Anwendungsszenarien für Suche und Exploration dargelegt.

Contents

Chapter 1

Introduction

1.1 Motivation

With the emergence of freely available knowledge bases there is an increased demand for analyses that span multiple knowledge and data repositories. Cross-lingual sentiment extraction, detecting inconsistencies over several knowledge bases and query answering on the Web are only some examples of analyses of this kind. The core of these intelligent systems is knowledge in a machine-readable representation: ontologies, taxonomies or link graphs. Collectively, we refer to them as *Web knowledge repositories* (KRs).

Depending on how formal a knowledge repository represents its content, we distinguish between ontologies, link graphs, taxonomies and dictionaries as illustrated in Figure 1.1. *Ontologies* represent one extreme - they serve as formal domain specifications. Studer [95] defines an ontology as "a formal, explicit specification of a shared conceptualization". Connections between elements are explicitly typed and called *relations*, as well as attributes are explicitly assigned to elements. The other extreme are *dictionaries*. They contain a plain list of items, without organizing them into a structure. *Link graphs and taxonomies* are in the middle of this spectrum and are widely used on the Web. Link graphs can be understood as a relaxed version of fully specified ontologies. Elements are interconnected; however, the meaning of the links is implicit and needs either human judgement or a machinery for type-casting them into relations. Taxonomies relax the link structure further. Individual items are solely connected to the parent category.

Originating from different communities or enterprises knowledge repositories are highly diverse and disconnected. To jointly use disconnected repositories in an application, there is a need for finding correspondences between their objects

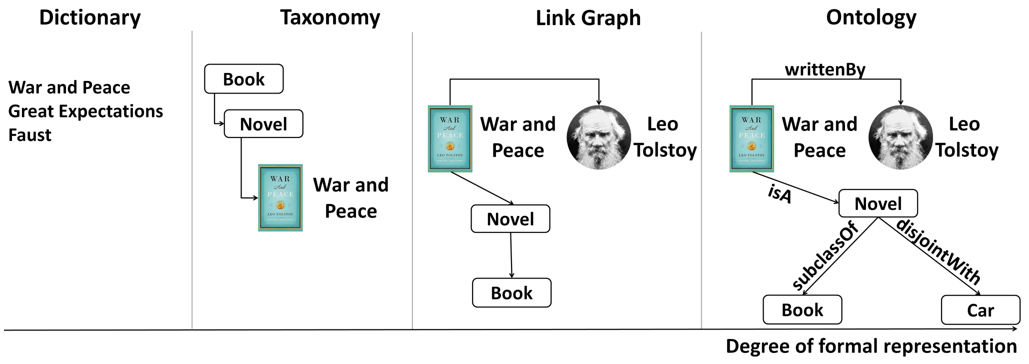


Figure 1.1: **Different kinds of knowledge repositories organized by their degree of formal representation.**

- categories and instances. This is the task of *knowledge repository alignment*. Alignment methods operate on element and structure levels. On the element level, local information about an element is harnessed to infer matches. Title or attribute values of an element are examples of local element information. On the structural level, alignment methods use structural patterns to detect matching objects.

As illustrated in Figure 1.1, depending on the representation forms different level of information is provided for elements and structures. Ontology matching research addresses the problem of finding an alignment between two ontologies. Many of the proposed methods rely on the sufficient overlap on both, element and structure levels. The Linked Open Data (LOD) community has developed guidelines for incorporating disjoint ontologies into a common cloud. To become a part of the interlinked cloud, a data set has to be exposed in RDF format, use URIs to index objects and reference objects that are already in the cloud. In 2014, more than one thousand of data sources were integrated into the LOD cloud.

The problem of finding alignment between dictionary entries arise in computational linguistics when detecting correspondences between words or multi-word phrases. Usually, dictionaries do not contain any structure, and methods for their alignment heavily rely on background knowledge to infer additional meaning of the entries or to provide the dictionary with a structure.

The focus of this dissertation is on the middle-ground between these two extremes - on Web link graphs and taxonomies. Being less expressive than ontologies, knowledge represented as taxonomies or link graphs compose a considerable amount of Web resources. Taxonomies, for instance, are the underlying models of Web directories or product catalogs of online marketplaces (such as icecat.biz and amazon.com). According to Eurostat[1], there are 1.6 million enterprises in the EU. 20% of them conduct e-sales organizing their goods as catalogs. One of the most prominent examples of link graph is Wikipedia and its underlying MediaWiki[2] engine. MediaWiki is an open source software, which can be deployed for private and enterprise use. Apart from MediaWiki, there are other wiki engines such as DokuWiki[3] or PmWiki[4]. There are at least 10,000 wiki-based Web sites[5], covering different domains from books and movies to recepies. These numbers illustrate the scale at which taxonomies and link graphs are presented on the Web and the need in methods for producing alignments over them. In the next section we review the progress made to date in aligning link graphs and taxonomies.

1.2 State-Of-the-Art

Finding matches between entries of two data sets has been addressed in previous research in several domains.

Schema matching applications [21, 24, 62] aim at finding equivalent elements of two schemas - relational databases or XML schemas. The found matches are used subsequently for constructing a global view of several local schemas. On the structure level, matchers operate with table and attribute names, existing unique, primary and foreign keys, attribute data types and relationship cardinalities. When the schema information is limited or there are several equally possible matches, analysing linguistic patterns of the corresponding attributes or their value ranges [2, 56, 99] may help to make a matching decision.

Providing unified access to heterogeneous data sources is the overriding goal of the field of **data integration** [41, 53, 103]. A global mediation schema is exposed to end-users or applications for querying. The key task is to find a mapping between the mediation schema and the local schemas of the underlying sources. Data integration applications require (a) the mediation schema to

[1] `http://ec.europa.eu/eurostat/statistics-explained/index.php/E-commerce_statistics`, accessed on 27.06.2017

[2] `https://www.mediawiki.org/wiki/MediaWiki`, accessed on 27.06.2017

[3] `https://www.dokuwiki.org/dokuwiki#`, accessed on 14.06.2017

[4] `http://www.pmwiki.org/`, accessed on 14.06.2017

[5] `http://s23.org/wikistats/largest_html.php`, accessed on 13.02.2017

maintain up-to-date mappings to local schemas; (b) mechanisms or integrating many local sources without loading them into the central warehouse. On the Web, these principles were adopted by booking or retailer site aggregators.

For both schema matching and data integration approaches, the rich structure of knowledge repositories is a key to producing matchings. Typically, the matchings are found at the structure level - find correspondences between columns of database tables or between elements of Web forms. Producing matchings for individual records in the repositories are usually considered out of the scope.

Numerous online marketplaces enable merchants selling their products and process transactions via a centralized marketplace provider. **Catalog integration** applications [1, 46, 79] offer mechanisms for (semi-)automatic loading merchant products into the product catalog of an online marketplace. Each merchant has to ensure the integration of his products onto the global classification provided by the online marketplace. Usually, merchants already have some product categorization which can be used for establishing integration rules such as "if two products are in the same category in merchant catalog, they may not be placed to marketplace categories distant from each other". This type of alignments is of "instance-to-category" type. Establishing explicit alignments between categories of both catalogs is usually neglected. Another limitations of catalog integration approaches is relying on matching instances between merchant and marketplace catalogs, which can not be guaranteed in case of KRs acquired Web-wide.

The field of **ontology alignment** typically considers logically rigorous ontologies like OWL or RDF schemas along with the instances of classes and properties [93]. There is a wealth of prior work on ontology alignment in this spirit; representatives and overviews include [30, 36, 96, 102]. The Ontology Alignment Evaluation Initiative[6] provides test cases of two kinds: different ontologies translated into different languages and same ontologies translated into different languages. Most of the existing approaches are applicable for small- and medium-size ontologies. Their run time performance is pushed to the limit when the alignments have to be computed at scale.

[6] `http://oaei.ontologymatching.org/` accessed on 14.03.2017

1.3 Challenges

The goal of this dissertation is to study methods for constructing and curating an alignment between Web KRs. Given shortcomings of the state-of-the-art solutions, this requires addressing the following challenges.

Terminological Heterogeneity

State-of-the-art approaches in ontology alignment or schema matching take the diversity of the data sets as a call for full-fledged data integration or top-down standardization. This is an infeasible solution when the knowledge repositories are enormously diverse and simply do not share equivalent categories or instances. To illustrate the shortcoming of full integration approach, consider the book category *Kinder- & Jugendliteratur* on amazon.de (en. Children & Youth Literature) and two relevant counterparts on amazon.com - *Children's Books* and *Teen & Young Adult*. Both counterpart categories are highly relevant, not matching the name of the source category perfectly, though. It signals the importance of finding alignments that allow users to navigate across the boundaries of knowledge repository and explore different repositories together, while living with existing terminological diversity.

Structural Heterogeneity

On the Web, a large number of knowledge repositories specifically focus on a single application and community. This is opposite to the definition of ontologies, which are "shared conceptualization" and have "to be used and reused across different applications and communities" [104]. As a result, taxonomies and link graphs of KRs have highly diverse structures, which are challenging to deal with for the state-of-the-art methods. This requires to go beyond finding perfectly matching categories or instances in both KRs, since their overlap is very small. For example, when aligning Wikipedia with Eurostat's Wiki a full integration approach is not able to discover a potentially interesting alignment between pages *Bioethanol* and *Südzucker* (a German producer of biofuel), since Eurostat's Wiki comprises solely general concepts and does not hold organizations, people or locations. This motivates the demand for approaches, which are able to perform alignment given highly diverse structures of two taxonomies or link graphs.

Dynamic Repositories

Web knowledge repositories such as Wikipedia have a large number of contributing editors and keep growing at impressive rates. Conventional ontology alignment and data integration tools rely on static sources - they consider snapshots of datasets while interlinking. For large data sets, recomputing alignments from scratch every time an update occurs in one of the sources is expensive. An alignment between two knowledge repositories can be used for filling one repository with the information from the aligned one. Consider constructing a link graph between instances of different type like people or organizations from a news feed. Once a connection between two instances appears in the news (prominent people get married or take a new office), the constructed link graph has to include this connection. Ideally, with a minor delay. This motivates the need in supporting an efficient procedure of alignment updates and knowledge repositories maintenance for large and dynamic KRs.

1.4 Contribution

This dissertation addresses the challenges outlined in the previous section. In particular we make the following contributions.

ACROSS: Reducing Terminological Heterogeneity

The first contribution of this dissertation is the ACROSS system for supporting alignment of knowledge repositories with high terminological diversity. The terminology of a KR is formed by its unique culture, which includes language and socio-economic background. To resolve terminological heterogeneity, we harness Wikipedia as background knowledge. This enables ACROSS to find category-category matches between two knowledge repositories of highly diverse and potentially multilingual classification systems. Additionally, ACROSS performs constraint-aware reasoning to ensure consistent alignments. ACROSS's results are published in:

- [83]: N. Prytkova, M. Spaniol, and G. Weikum. Aligning Multi-Cultural Knowledge Taxonomies by Combinatorial Optimization. In *Proceedings of the 24th ACM International Conference on World Wide Web*, WWW 2015.

- [11]: N. Boldyrev, M. Spaniol, and G. Weikum. ACROSS: A Framework for Multi-Cultural Interlinking of Web Taxonomies. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci 2016.

Additionally, this work is under peer review:

- [12]: N. Boldyrev, M. Spaniol, and G. Weikum. Multi-Cultural Interlinking of Web Taxonomies with ACROSS. *The Journal of Web Science*, under review.

LILIANA and SESAME: Reducing Structural Heterogeneity

The second contribution of this dissertation are the LILIANA and SESAME systems. They target alignment of knowledge repositories created by different communities. Detecting related instances in the counterpart knowledge repository given low structural similarity is the focus of both systems. We harness link structures of knowledge repositories to assess the similarity between instances. As a use-case, LILIANA reconstructs an alignment between YAGO type taxonomy and the category hierarchy of Eurostat Wiki. This alignment is a key to discovery relevant Eurostat statistics for textual media like news. In contrast to LILIANA, SESAME's focus is on link graphs rather than taxonomies. Consider two link graphs - Wikipedia and Eurostat Wiki. There is a very small overlap of perfectly matching instances. These, however, are generic concepts - *Biofuel* or *Unemployment* to name some. To unlock the rest of Wikipedia link graph, which is highly related to the perfectly matching general concepts, is SESAME's goal. This work is published in:

1. [89]: M. Spaniol, N. Prytkova, and G. Weikum. Knowledge Linking for Online Statistics. In *Proceedings of the 59th World Statistics Congress*, WSC 2013.
2. [10]: N. Boldyrev, M. Spaniol, J. Strötgen, and G. Weikum. SESAME: European Statistics Explored via Semantic Alignment onto Wikipedia. In *Proceedings of the 26th ACM International Conference on World Wide Web Companion*, WWW 2017.

LAIKA: Supporting Alignment of Dynamic Knowledge Repositories

The third contribution of this dissertation is the LAIKA system for supporting interlinking of large and dynamic graphs. Each Wikipedia edition is an autonomously curated knowledge repository with manually created inter-language links to other language editions. The interlinking of categories and articles across language editions of Wikipedia are to be understood as an alignment. Wikipedia content is highly dynamic, capturing information about recent events and prominent people. However, for events that are of regional interest and for long-tail concepts and categories, there is often a big delay before they are picked up by the community and get appropriate inter-language linking. We utilize the extensive Wikipedia link graph for recommending for a given category inter-language links. Additionally, the link structure of one Wikipedia edition is exploited to recommend missing article-article and article-category links for the aligned counterpart Wikipedia. LAIKA's results are published in:

1. [37]: J. Göbölös-Szabó, N. Prytkova, M. Spaniol, and G. Weikum. Cross-Lingual Data Quality for Knowledge Base Acceleration across Wikipedia Editions. In *Proceedings of the 10th International Workshop on Quality in Databases*, QDB 2012.

1.5 Outline

The rest of this dissertation is organized as follows. *Chapter 2* gives definitions of ontologies, link graphs, taxonomies and alignment; reviews methods for dealing with terminological and structural heterogeneities; outlines approaches to alignment curation and ends with summary of design choices for alignment pipelines. *Chapter 3* presents the ACROSS system for reducing terminological heterogeneity. We address the problem of finding alignment between categories of two taxonomies via obtaining semantic labels from Wikipedia and using structural constraints to produce concise and coherent alignments. *Chapter 4* gives an overview of LILIANA and SESAME, two systems for bridging the gap between KRs with highly heterogeneous structures. *Chapter 5* is dedicated to the LAIKA system, which focuses on curation alignment between two large-scale and highly dynamic Web link graphs. Additionaly, we describe how aligned link graphs can be maintained - in particular, how missing article-article links can be detected. We conclude in *Chapter 6* by revising three problems addressed in this dissertation, discuss our contributions and give an outlook of possible research directions.

Chapter 2

Foundations and Technical Background

This chapter places our contributions into the context of previous work and gives background with emphasis on methods for resolving terminological and structural heterogeneity for aligning Web knowledge repositories (KRs). We begin with formal definitions of link graphs, taxonomies and ontologies. Next, we review the tasks of alignment construction and curation, as well as knowledge repository maintenance. We conclude with reviewing different models for composing alignment pipelines.

2.1 Knowledge Repositories

In this section we review knowledge repositories having different degree of formal representation along with their applications. Throughout this dissertation, we assume that a knowledge repository is either a taxonomy or a link graph. For completeness, we also give an overview of ontologies.

Next we outline some conventions used in the reminder of this dissertation.

Definition 2.1. A *knowledge repository* (KR) consists of instances I and classes C. Instances refer to individual entities such as books or drugs. A class (or category) is a group of instances sharing a common property. We collectively refer to instances and categories of a KR as *KR objects*.

In the following subsections, we consider taxonomies, link graphs and ontologies in detail.

2.1.1 Taxonomies

The focus of taxonomy-based applications is on building hierarchical indexing of their contents to enable browsing over large corpora.

For example, online library categorization systems (the Dewey Decimal Classification used by the German National Library[1], the Library of Congress Classification[2]) offer browsing library contents via a hierarchy of topics. The Anatomical Therapeutic Chemical (ATC) classification[3], the International Statistical Classification of Diseases[4] and the Medical Subject Headings (MeSH) classification[5] serve as reference points for drugs, diseases or indexing medical publications. Each instance of these classification systems - a book, an active ingredient or a disease - is referenced from the category system. Co-occurring of instances within the same category and distance in the category tree can be used as a proxy for measuring similarity between instances. Fokoue et al. [33] incorporate ATC classification for predicting drug-drug interactions. They compute similarity between the paths in the ATC hierarchy that correspond to the drugs.

Definition 2.2. A *knowledge taxonomy* is a directed acyclic graph (DAG) with nodes of two types: instances and categories (referred to as I and C respectively). Depending on the type of nodes they connect, we distinguish between: *i-c* edges for connecting instances to their categories and *c-c* edges for connecting parent category to its child category.

2.1.2 Link Graphs

Graphs find their applications in many domains as a simple yet powerful model for expressing interaction between data elements. In contrast to taxonomies, link graphs provide interlinking over instances as well. To illustrate the wide variety of Web data which is usually modelled as link graphs, we review several prominent examples.

[1] `http://deweysearchde.pansoft.de/webdeweysearch/mainClasses.html?catalogs=DNB` accessed on 27.06.2017

[2] `https://catalog.loc.gov/vwebv/searchBrowse` accessed on 27.06.2017

[3] `https://www.whocc.no/atc/structure_and_principles/` accessed on 27.06.2017

[4] `http://apps.who.int/classifications/icd10/browse/2016/en` accessed on 27.06.2017

[5] `https://meshb.nlm.nih.gov/#/treeSearch` accessed on 27.06.2017

ACM digital library[6], PubMed[7] (a collection of biomedical articles and books), CiteSeerX[8] or Google Scholar[9] – are typical examples of citation graphs with nodes being publications and each reference to another publication being a directed edge. Some of the publication collections refine the graph structure by expressing each author as a node and adding edges between an author and his publications as well as edges to all his co-authors. The "networked information space" [61] of this kind enables detecting most co-cited publications, most authoritative works and authors to enrich search and navigation functionalities of the electronic databases.

Social Web sites like Twitter[10] or Instagram[11] are directed graphs of users and their *friendship*, *repost*, *follow* or other relations. Bronson et al. [14] describe the Facebook[12] social graph model. In addition to the user nodes, it includes object nodes such as locations and comments. The edges are annotated with the time stamp and encode actions between the nodes, e.g., *commented* or *tagged*. Social network analysis is a big research topic, which aims at, among others, describing properties of social networks and detecting communities as studied by Yang et al. [109] and Metzler et al. [65] or modelling memes spread as proposed by Leskovec et al. [54].

Finally, the Web by itself is usually represented as a graph. Each Web page is a node and hyperlinks between pages are edges. The tasks of document retrieval [78] and link farms detection [107], for instance, rely on the link structure of the Web.

In this dissertation, we focus on a special type of Web link graphs called Wiki link graphs. Wiki-based knowledge repositories such as Wikipedia, Wikiquotes and other public and private knowledge management systems built with Wiki engine have several properties in common. All of them have a backbone categorization system, where each category is a Web page. A category page differs from an article page in the way it is structured and used. A category page does not hold any content related to the topic and it exclusively serves categorization and navigation purposes. In contrast, an article (or a content page) is an encyclopedic-style entry about a subject. Following WikiMedia guidelines, an article links to other related articles and should be reached trough Wiki

[6] `http://dl.acm.org/` accessed on 27.06.2017

[7] `https://www.ncbi.nlm.nih.gov/pubmed` accessed on 27.06.2017

[8] `http://citeseerx.ist.psu.edu/index` accessed on 27.06.2017

[9] `https://scholar.google.de/` accessed on 27.06.2017

[10] `https://twitter.com/` accessed on 27.06.2017

[11] `https://www.instagram.com/` accessed on 27.06.2017

[12] `https://www.facebook.com/` accessed on 27.06.2017

category system[13] [14]. Since all Wiki-based KRs are collaboratively modified, the underlying link graph is highly dynamic.

Throughout the thesis, we refer to Wiki links graphs as link graphs for conciseness.

Definition 2.3. A *link graph* is a directed graph composed of nodes of two types: instance and category nodes (referred to as I and C respectively). There are directed edges between nodes. *Subcategory-of* edges are defined over category nodes, and instances are connected to categories with *instance-of* edges (*i-c* and *c-c* edges). One instance references another instance via an *i-i* edge. In case of Web link graph, *i-i* edges are links between documents.

2.1.3 Ontologies

Ontologies capture extended knowledge about a domain and formalize this knowledge by:

1. type-casting links between instances or classes into relations;
2. defining attributes and their types for instances;
3. fixing domain and range types for relations; and
4. providing a set of axioms for reasoning about instances and relations.

Each assertion in an ontology is a fact and it can be represented as a triple of $\langle subject, \; relation, \; object \rangle$. *Subject* and *object* are to be understood as nodes in a graph and *relation* is a typed edge between them.

Definition 2.4. An *ontology* is a directed graph composed of nodes of several types: instances I, categories C and attributes V. There is a set of relations R which define the types of the edges in the ontology. For example, an ontology might include *marriedTo* or *capitalOf* relations.

Recent advances in large scale information extraction [70] and automatic ontology construction [19] gave rise to endeavours in academic world such as YAGO [97], DBpedia [4], and KnowItAll [29], as well as in commerce – Freebase [13] which is a part of Google Knowledge Graph [87]. These are examples of

[13] `https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article` accessed on 14.03.2017

[14] `https://en.wikipedia.org/wiki/Category:Articles` accessed on 14.03.2017

general-purpose ontologies. UMLS [9] or KnowLife [28] are tailored to the medical domain. They cover diseases, drugs, active ingredients and other types of instances, as well as domain-specific relations.

Ontologies became a core part of many text understanding applications. Named entity recognition and disambiguation (NERD) is a prominent task of that kind. Mentions of entities such as people, organization or locations are detected in free text. Subsequently, each mention is mapped to the most similar entity registered in an ontology. This type of text annotations can be understood as a semantic indexing and used, for instance, for semantic-enriched browsing over a news collection [43].

2.2 Alignment Construction and Curation

In order to perform analysis spanning several knowledge repositories the correspondences between instances and categories belonging to different KRs have to be found.

Depending on the application, there are many types of relations that can hold between corresponding objects of two knowledge repositories. In ontology matching research [30] alignment methods might target, for instance, *less general* or *equivalent* relations between classes.

Given the inherent heterogeneity of Web knowledge repositories, we aim at discovering *related-to* correspondences between objects of two KRs. We align objects of the same type: instances are aligned to instances in a counterpart knowledge repository and categories to categories. The instance-to-category type of alignment is addressed in the catalog integration research by placing products into correct categories in a catalog. Discovering this type of correspondences is beyond the scope of this dissertation.

With each correspondence we associate a confidence value - a measure of certainty that two objects are indeed related. Throughout the dissertation we assume the following definition of alignment.

Definition 2.5. An *alignment* $\mathcal{A}$ between two knowledge repositories K_1 and K_2 is a set of triples $\langle o_1, o_2, c_{12} \rangle$, where either $o_1 \in I(K_1)$ and $o_2 \in I(K_2)$ or $o_1 \in C(K_1)$ and $o_2 \in C(K_2)$. $c_{12} \in \mathbb{R}$ is a confidence value associated with the alignment between o_1 and o_2.

Given this definition, we consider the problems of alignment construction and alignment curation.

The task of **alignment construction** is to find a set of correspondences $\mathcal{A}$ given two Web knowledge repositories K_1 and K_2.

In the field of schema matching, most alignment methods target $1-1$ matches (i.e., discovering equivalence relations between tables and attributes) by mapping to a counterpart with the highest confidence score. For Web knowledge repositories, pursuing $1-1$ mapping is a challenging task. As an example, the book category *Kinder- & Jugendliteratur* on amazon.de (en. Children & Youth Literature) has two relevant counterparts on amazon.com – *Children's Books* and *Teen & Young Adult*, not matching them perfectly, though. In this dissertation, we assume $n-n$ alignment cardinality, which means that for each instance or category in a knowledge repository there is a list of possible counterparts ranked by alignment confidence.

However, an alignment might have a more complex structure rather than a ranked list of counterparts. The iMap system developed by Dhamankar et al. [21] targets finding mapping rules between the attributes of two databases. They can have a complex structure such as:

$$\begin{aligned} name &= concat(first\_name, last\_name) \\ listed\_price &= price * (1 - tax\_rate) \end{aligned} \tag{2.1}$$

Existing keys, constrains, and query logs are the vital assets that support finding mapping rules between tables of relational databases. Despite the progress made up-to-date in data and schema integration research, complex alignment rules of this kind are still challenging to infer in general [23, 73]. Discovering compositional alignments is out of the scope of this dissertation.

As outlined in the Introduction, Web knowledge repositories are highly dynamic[15] . Figure 2.1 illustrates the growth of Wikipedia between the years 2001 and 2008. Starting from 2005, the average number of added articles per day was above 1000.

Once an alignment between two knowledge repositories is constructed, it needs to be maintained to ensure its validity and completeness over time. This is the **alignment curation** task.

The problem of ensuring an up-to-date mapping is highly relevant to data integration applications. Consider product aggregation Web sites such as Trivago[16]

[15] `https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia` accessed on 27.06.2017

[16] `https://www.trivago.de/` accessed on 27.06.2017

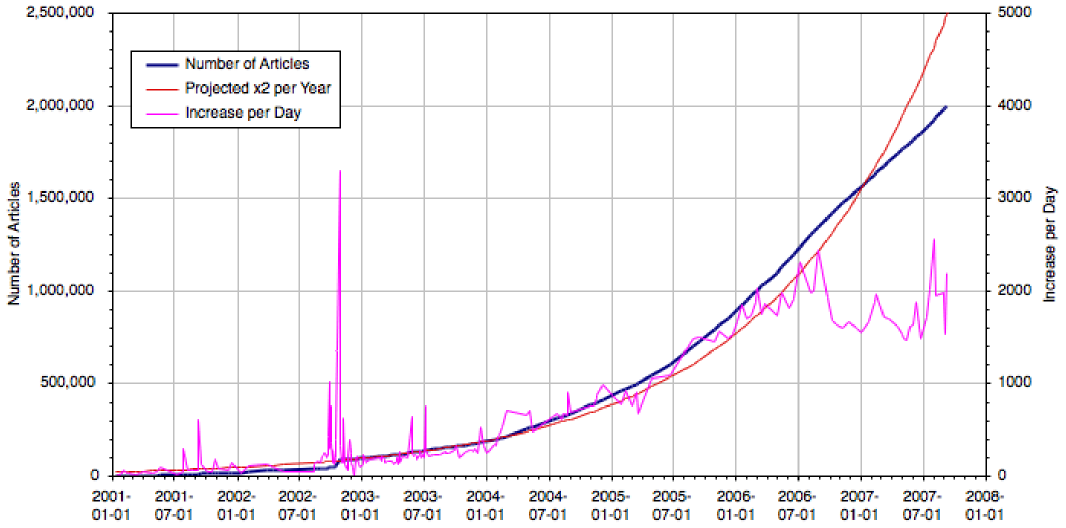


Figure 2.1: **Article count increase for en.wikipedia.org (purple line). The actual growth of Wikipedia is shown with the blue line. Its approximation based on the assumption of annual doubling using the article count on 01.01.2003 as the starting day is illustrated with the red line.**[19]

or Booking.com[17]. The schemas of local Web pages (individual product or service providers) might change and the global schema owned by the aggregation Web site has to keep track of local changes and maintain the mappings. Being successful at incorporating various transformation and mapping rules between schema elements, many applications still need to be rerun even when minor changes occur in integrated schemas. Running the mapping procedure from scratch creates considerable overhead, when integrated schemas or interlinked knowledge repositories are highly dynamic.

In data integration and data exchange applications, mappings between schemas are expressed in a declarative way, for example, as view definitions. When a schema evolves, the view definition (the mapping) has to be rewritten to reflect the new structure and semantics of the schema elements. For example, the ToMAS system developed by Velegrakis [105] automatically adapts mappings between dynamic schemas. ToMAS focuses on structural and logical associations between schemas that might change and invalidate current

[19] Source: ©User:Seattle Skier / Wikimedia Commons / CC-BY-SA-3.0, `https://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth` accessed on 21.04.2017

[17] `http://www.booking.com/` accessed on 27.06.2017

mapping. Consider adding a constraint that each record of table *Grants* is associated with exactly one record in the table *Funder*. The mapping rule *Grants.id* = *Project.id* between two databases is no longer conformed with the new structure, since transferring all records from table *Project* onto table *Grants* with the old mapping rule raises an exception. However, by appropriate adjusting the query, correct data exchange between two databases can be restored. ToMAS is tailored to relational tables and XML schemas and is inapplicable to link graphs and taxonomies, which are less formal.

2.3 Knowledge Repository Curation

Many knowledge repositories are constructed from an underlying corpus of documents. An example is the Open Directory Project [20], which organizes links to Web resources in a taxonomy of topics. An editor community is responsible for keeping the content up-to-date and add new links once a new high-quality resource is made available. Another prominent example is YAGO, a Wikipedia-based ontology. Currently, YAGO is constructed from a snapshot of Wikipedia. In the ideal case, new facts are made available in YAGO right after they are added to Wikipedia.

The task of **knowledge repository curation** is to update a Web knowledge repository K - add or update instances, classes and relations over them. Usually, curation tasks rely on external sources - collections of Web documents, news or knowledge bases.

Some of the tasks included by knowledge repository curation are:

- Completing a knowledge repository from an underlying document collection. NELL [15] (Never-Ending Language Learning) is a project whose goal is "to develop the world's largest structured knowledge base". Once equivalences between phrases in free text articles and relations in an ontology are established, one can infer new facts. For instance, the sentence

  "Nirvana is a legendary grunge band."

  produces the fact *bandGenre(Nirvana, Grunge)*, if the alignment $\langle X \textit{ is a legendary } Y \textit{ band}, \textit{bandGenre}\rangle$ is known.

- Detecting and quantifying incompleteness of a knowledge repository with respect to a reference collection of documents. Information extraction

[20] `http://dmoz.org` accessed on 14.03.2013

tools for building knowledge collections are constantly improving and are able to perform information harvesting with high precision. However, a document collection might not contain all the facts (e.g., might not name all the children of a person). Therefore, it is hard to judge how complete a knowledge repository is. Mirza et al. [68] use textual patterns to scan Wikipedia bibliographical articles and fetch the number of children for each person in Wikidata ontology. The fetched number is compared against the number of children known to Wikidata to estimate the recall of the *hasChild* relation.

- Reasoning with external tools and rules. Reasoning is a usual procedure in building knowledge repositories. It can target either cleaning existing data or inferring new facts by applying a set of rules. There are many formalisms to express rules. OWL (a language for ontology description) and logic programming are two prominent examples. A direct transfer of rules between two formalisms is not trivial since both rely on different assumptions. Logic programming assumes that all the existing facts are known. If a fact can not be proven, it is said to be false. OWL, in contrast, has open-word semantics and returns a fact which can not be proven as underspecified. To perform a hybrid reasoning - combining rules from different formalisms - an interface between two logic systems is needed. Some methods to narrow this gap are introduced by Drabent et al. in [26] and [25].

2.4 Terminological Heterogeneity

An alignment method can be seen as a function which, given two input knowledge repositories, produces a set of correspondences $\langle o_1, o_2, c_{12}\rangle$. To get the confidence value (or similarity) c_{12}, alignment methods harness node-level information as well as global structure of the knowledge repositories. At the end, both node- and structure-level similarities are aggregated to assign a single confidence value to a pair of potential counterparts. In this section, we focus on methods working with local information provided by nodes to take an alignment decision.

Almost all alignment methods in data integration, ontology alignment or schema matching rely on the information directly contained in an object. Depending on the level of formality, the amount of information varies between different knowledge representations. Link graphs and taxonomies annotate their nodes with titles. More formally represented objects such as attributes of relational tables or objects of an ontology have richer annotations (for example, data type of table attributes or set of key phrases for an ontological entity).

2.4.1 Intrinsic Techniques

Most basic similarity measures for assessing the similarity between two objects rely on the similarity of their string representation (e.g., titles). Typically, all linguistic matchers have a layered structure.

On the preprocessing layer, strings undergo lexical normalization such as case normalization, suppression of punctuation, digits and diacritics. Linguistic normalization methods are built upon more elaborate techniques. One of them is tokenization. Each string is split into words (tokens) according to splitting rules or delimiters. Stemming or lemmatization bring tokens to their root form or strip out inflectional endings and cast a token into its dictionary form. To this end, a sequence of tokens can be annotated with their POS tags to produce a lifted view of the string (e.g., by replacing articles with DT or adjectives with JJ).

On the similarity assessment level, the distance between two strings can be either string- or token-based.

One of the standard choices for string-based similarity is Levenshtein distance [55]. It is computed as a minimal number of change operations (deletions, insertions and character replacement) needed to transform one string into another.

Token-based methods are more robust to detecting resemblance between strings which contain roughly the same set of words. The input strings are split into tokens and the similarity is assessed at the token-level using any measure for set or vector similarity. Cosine similarity or Jaccard distance are some examples for the measures of that kind.

In summary, intrinsic methods measure similarity of two objects based on their labels.

2.4.2 Extrinsic Techniques

Intrinsic techniques are brought to their limits when knowledge repositories use different jargon, synonymic terms or are in different languages. Instead of directly comparing titles of two objects, extrinsic methods exploit external resources like lexicons or dictionaries (*mediators*) to infer semantics of a KR object. To this end, similarity between semantics of two objects is assessed.

Semantics of an object can be inferred from structured sources such as dictionaries or semantic nets of concepts. Prominent examples of structured mediators include WordNet [66], Wikipedia and MeSH (Medical Subject Headings)[21]. WordNet [66] is a semantic net of words and relations defined over them: hyponymy, meronymy (part-of), synonymy and their inverses. WordNet has been extensively harnessed in various applications such as for knowledge base construction described by Rebele et al. [84] or named entity recognition and disambiguation studied by Nguyen et al. [72].

Based on the information considered for similarity computation, we review taxonomy-based, synset-based, gloss-based and link graph-based similarity measures.

Taxonomy-based Similarity

The intuition behind taxonomy-based similarity is that two concepts are similar if they are close in a concept hierarchy. The more edges one has to traverse to arrive from one concept to another, the more dissimilar they are. Figure 2.2 illustrates a sample concept hierarchy. The concept *DOCTOR (MEDICAL)* is closer in meaning to *NURSE* than to *BUS DRIVER*.

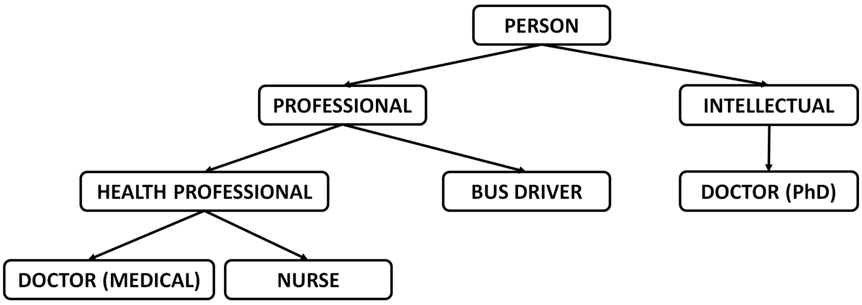


Figure 2.2: **An example of a concept hierarchy.**

Resnik [85] proposes to compute the similarity of two concepts c_1 and c_2 given a concept hierarchy as a distance to the least common hypernym:

$$sim(c_1, c_2) = \max_{c \in H(c_1,c_2)} -\log p(c) \tag{2.2}$$

with $H(c_1, c_2)$ denoting the set of common hypernyms.

Function $-\log p(c)$ describes information content of a concept c. It is 1 for the root node and decreases monotonically along the depth of the hierarchy.

[21] https://www.nlm.nih.gov/mesh/ accessed on 27.06.2017

However, in practice, object labels are highly ambiguous and can be mapped to several concepts. Consider the term *doctor* which can be mapped both to *DOCTOR (MEDICAL)* and *DOCTOR (PhD)*. Resnik's formula for measuring similarity between two labels l_1 and l_2 is:

$$sim(l_1, l_2) = \max_{\substack{c_1 \in concepts(l_1) \\ c_2 \in concepts(l_2)}} sim(c_1, c_2) \tag{2.3}$$

where $concepts(l_1)$ are all possible concepts for the label l_1. $concepts(l_2)$ has similar meaning.

Taxonomy-based metrics are, however, highly sensitive to the classification criteria used for constructing the underlying classification of concepts.

Synset-based Similarity

WordNet maps each term to a set of its synonyms (synsets). For example, *doctor* is synonymic to *doctor, doc, physician, MD, Dr., medico.*

Euzenat et al. [30] describe similarity of two concepts as Jaccard similarity of their synsets:

$$sim(c_1, c_2) = \frac{|S(c_1) \cap S(c_2)|}{|S(c_1) \cup S(c21)|} \tag{2.4}$$

where $S(c_1)$ and $S(c_2)$ are the synsets for concepts c_1 and c_2.

Gloss-based Similarity

In dictionaries and semantic word networks, a word is also annotated with a verbal description of its meaning (a *gloss*). For example, *doctor: a licensed medical practitioner.* Banerjee et al. [6] use similarity of concept glosses as a proxy to concept similarity:

$$sim(c_1, c_2) = \sum |overlapping\_sequence(gloss(c_1), gloss(c_2))|^2 \tag{2.5}$$

where *overlapping_sequence* is an overlap of consecutive content words. This formula explicitly prefers encountering phrase overlap of n words long rather than n overlaps of length one.

To capture more context for a concept, its gloss can be extended by the glosses of its hypernyms or hyponyms as shown by Banerjee et al. in [7].

Aggregated similarity is composed of gloss similarities of all possible pairs of concepts related to c_1 and c_2:

$$sim_{ext}(c_1, c_2) = \sum_{\forall R_1, R_2} sim(R_1(c_1), R_2(c_2)) \tag{2.6}$$

where $R_1(c_1)$ is a concept in the WordNet graph connected to c_1 via relation R_1 (for example, *hypernym* or *hyponym*). R_2 has analogous meaning.

Instead of annotating a concept with its WordNet gloss, some approaches (e.g., developed by Gracia et al. [39]) annotate concepts with descriptions from Web data. Paulheim [80] proposes fetching concept descriptions from results of a search engine. Each concept is supplied to a search engine as a query and snippets of the returned relevant pages are concatenated into one large concept description.

Similarity Based on Link Graph

Large interlinked knowledge collections such as Wikipedia contain a large number of cross-referenced pages. Pages belonging to the same topic (like pages about ecology) tend to be better interlinked than pages about distant topics (e.g., *War and Peace* and *Automobile*). Milne et al. [67] quantify similarity between two concept pages by the overlap of their hyperlinks:

$$sim(c_1, c_2) = \frac{\log(\max(I(c_1), I(c_2)) - \log(|I(c_1) \cap I(c_2)|))}{\log(W) - \log(\min(I(c_1), I(c_2))} \tag{2.7}$$

where W is the total number of pages in Wikipedia and $I(c_1)$ denote the set of incoming links for concept page c_1.

Turdakov and Velikov [101] additionally define weights for various types of links when computing Wikipedia-based similarity. Thus, two pages connected via a *See also* link are considered to be highly related, whereas links to date pages have the lowest weight.

Multilingual Similarity

A special case of terminological heterogeneity is multilingualism - comparing category and instance titles which are in different languages.

Usually, a pivot language is selected and titles of knowledge repositories are converted to it with the help of a dictionary or a translation service. Spohr et al. [92] and Fu et al. [34] refer to this scenario as cross-lingual matching. Let KR_1 and KR_2 be two knowledge repositories to be aligned. They expose titles of their objects in languages l_1 and l_2 respectively. Possible similarity computation strategies are:

- fix l_2 as the pivot language, translate all titles in KR_1 to l_2 and perform similarity computation in language l_2;
- same as above, but l_1 is fixed as the pivot;
- choose a third language $l_3 \neq l_2 \neq l_1$, translate both KR_1 and KR_2 to l_3 and compute title similarities.

Some knowledge repositories such as YAGO already provide titles in several languages for its categories, instances and relations. That is, each object in KR_1 is annotated with titles in a set of languages L_1 and in KR_2 in languages L_2. Then, to compute similarity between two objects in KR_1 and KR_2, one can compare titles of these objects in each of the languages $L_1 \cap L_2$. In this scenario, no mediator is needed and this can be considered as an intrinsic technique.

2.5 Structural Heterogeneity

The similarity methods described above use node-level information in isolation to make an alignment decision. In some cases, however, making a correct decision is not possible without having a global view on the knowledge repository. Take the book category *Europe* in a product catalog as an example. Without knowing its context (its place in the hierarchy and books belonging to it) one can not unambiguously detect its meaning - whether it is a category containing travel guides or books about European history. In this section we review methods working with structure-level information.

2.5.1 Graph-based Techniques

Alignment methods based on link structures of knowledge repositories assume that similar nodes must appear with similar link structures.

Anchor-PROMPT is a semi-automatic ontology alignment tool developed by Noy and Musen [77], which heavily relies on the link structure. The system expects a user input - a set of correct alignments (anchors). The anchors are used to guide the alignment procedure towards correct matching decisions. Discovering further alignments is done relying solely on the graph structure of input ontologies. We illustrate the algorithm by the example adopted from [77] (cf.

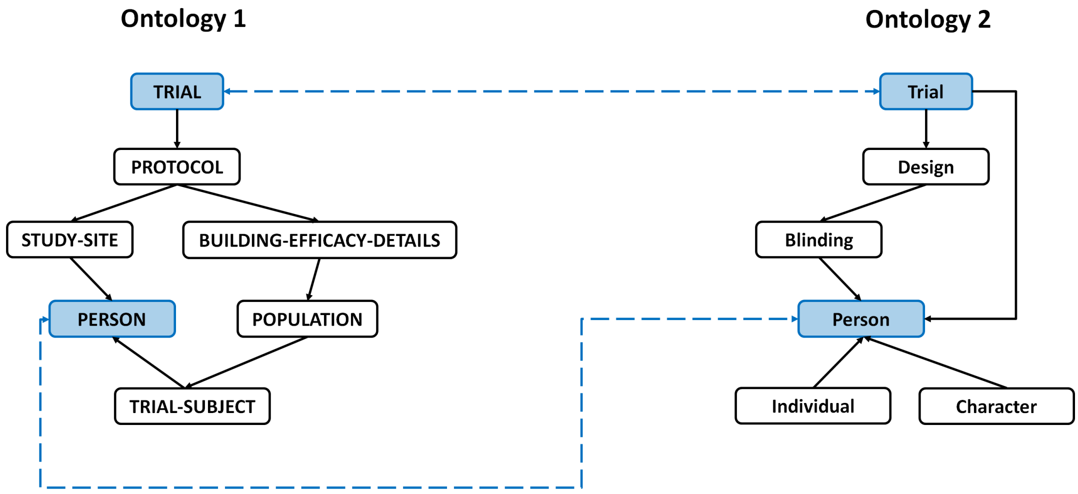


Figure 2.3: **An example of two ontologies with a partial alignment. The black arrows are the relations between the concepts. For conciseness, relation labels are omitted. The concepts marked with blue color are the starting and the terminal positions. Blue arrows denote the correct alignment provided by the user.**

Figure 2.3). The blue nodes represent the alignments provided by the user. $TRIAL$ and $trial$ are taken as starting concepts, and $PERSON$ and $person$ as terminal concepts. Anchor-PROMPT considers all possible paths between the starting and the terminal concepts in each ontology. Every time a pair of concepts appears in the same position in both paths, its similarity increases by a predefined value α. Assume, the two paths are generated:
$TRIAL \rightarrow PROTOCOL \rightarrow STUDY-SITE \rightarrow PERSON$ and
$trial \rightarrow design \rightarrow blinding \rightarrow person$.
Then, $sim(PROTOCOL, design)$ and $sim(STUDY{-}SITE, blinding)$ increase by α. Obviously, the more similar the link structure around two concepts is, the more likely they appear in the same position in the paths generated for each ontology.

Aligning two link graphs is similar in spirit to finding matching subgraphs. Zampelli et al. [110] proposed a method for approximate constrained subgraph matching. A classical subgraph matching problem is: given a pattern graph $G_p = (V_p, E_p)$ and a target graph $G_t = (V_t, E_t)$, find a mapping function $f : V_p \rightarrow V_t$ such that $(v_1, v_2) \in E_p \Leftrightarrow (f(v_1), f(v_2)) \in E_t$. For an approximate matching, one can design a set of constraints such as optional nodes or forbidden edges. Ontologies, for example, have rich information on ranges and domains for relations, which can be easily converted into a set of forbidden edges.

Another system for finding a fuzzy match of a pattern graph in a target graph is SAGA developed by Tian et al. [100]. Instead of predefining forbidden edges and optional nodes, SAGA devises a subgraph distance function which penalizes a matching between two subgraphs if:

- two subgraphs are structurally different;
- matched nodes have different labels, and
- there are nodes in the pattern graph, which cannot be matched to any nodes in the target graph.

Among all possible subgraphs in the target graph, SAGA selects the one with the minimal distance to the pattern graph.

2.5.2 Taxonomy-based Techniques

The taxonomic structure of a knowledge repository – its *subcategoryOf* hierarchy – can be exploited in the structure-level alignment methods in two ways. On one hand, structural similarity of two subtrees can act as a proxy to the semantic similarity of their roots. On the other hand, taxonomic structure can be used for detecting and repairing inconsistent alignments.

Spohr, Hollink and Cimiano [92] describe a method for aligning two taxonomies of financial accounting standards (FAS). These taxonomies have a specific structure. In addition to a hierarchical structure of financial terms, a FAS also provides an information on how a particular financial concept is calculated. For example, $assets = current\ assets + non - current\ assets$, where *assets*, *current assets* and *non-current assets* are concepts in a FAS taxonomy. In turn, each term also can be a combination of several concepts. This results in a hierarchy of computations of financial concepts. The authors propose a learning method to aligning two taxonomies, defining a variety of structure-based features. Let c_1 and c_2 be two concepts and $C(c_1)$ and $C(c_2)$ be their composing subconcepts. Assume that the calculation of a financial concept is preformed in a similar manner across countries. Then, c_1 and c_2 must have the same number of composing subconcepts. Formally, the taxonomy-based similarity of c_1 and c_2 can be expressed as:

$$sim(c_1, c_2) = 1 - \frac{||C(c_1)| - |C(c_2)||}{\max(|C(c_1)|, |C(c_2)|)} \tag{2.8}$$

The impact of structural features on increasing alignment performance has been shown experimentally by Do and Rahm [22] and Madhavan et al. [63]. However, the above formula is quite restrictive and can not be transferred to the cases when categorization criteria differs substantially across KRs.

Noy and Musen [76] introduce PROMPT tool for ontology merging with an incorporated repair module for keeping alignments coherent with respect to the taxonomic structures of source ontologies. PROMPT automatically detects as many equivalent categories as possible based on category names. This initial alignment is presented to the user, who can take one of the pre-specified actions - merge aligned categories, make a deep copy of a category from one ontology (copying all child categories as well), or make a shallow copy by taking over only the category itself. Each of the actions might potentially lead to an incoherent alignment or to discovering more alignments. Assume two categories c_1 and c_2 from source ontologies are fused by a user into a new category c in the merged ontology. Some of the actions which PROMPT can take relying on the category systems of both ontologies are:

- alert a user if c has more than one path to one of its parents;
- suggest connecting parent categories of c_1 and c_2 to c with $subclassOf$ relation; and
- discover all pairs of linguistically similar child categories in case c_1 and c_2 are marked by a user as equivalent.

The performance of PROMPT is highly dependent on the initial guess it makes to produce equivalent categories. If terminological overlap between two ontologies is poor, an extensive merging can not be guaranteed.

2.5.3 Instance-based Techniques

A strong signal of category equivalence can be obtained from the instances that the categories share. Ichise et al. [46] propose using k-statistics developed by Fleiss [31] to detect pairs of equivalent categories between two taxonomies. For each pair of categories, two probabilities are calculated. P denotes the probability of categories c_1 and c_2 being equivalent due to expressing the same concept and P' is the probability of c_1 and c_2 being equivalent by chance:

$$P(c_1, c_2) = \frac{N_{11} + N_{22}}{N_{11} + N_{12} + N_{21} + N_{22}} \tag{2.9}$$

$$P'(c_1, c_2) = \frac{(N_{11} + N_{12})(N_{11} + N_{21}) + (N_{21} + N_{22})(N_{12} + N_{22})}{(N_{11} + N_{12} + N_{21} + N_{22})^2} \tag{2.10}$$

Symbols N denote the values in the contingency table for categories c_1 and c_2 (Table 2.1).

	Category c_1	
Category c_2	belong	not belong
belong	N_{11}	N_{12}
not belong	N_{21}	N_{22}

Table 2.1: **Contingency table for sharing instances between two categories c_1 and c_2.**

The value of the k-statistics is computed as $k = \frac{P-P'}{1-P'}$.

Ichise et al. [46] propose to compare categories of two taxonomies in a top-down fashion and entirely rely on the assumption, that similar categories comprise similar sub-categories. First, using k-statistics equivalences between the root categories of the taxonomies are induced. To shrink the search space, subtrees rooted at the equivalent categories are traversed recursively to compute k-statistics between all possible category pairs between two taxonomies.

The drawbacks associated with bounding the exploration space by subtrees rooted at the already aligned categories were discussed in the previous section. Additionally, to produce an alignment covering the majority of categories in both taxonomies, one has to supply an extensive set of resembling instances. For some use cases, resembling instances can be detected automatically, for example, by finding matching URLs in two Web directories. However, if both taxonomies do not share instances or equivalent instances can not be easily detected, a sophisticated matching procedure or manual labelling has to take place.

2.6 Alignment Pipeline Models

The techniques described in the last two sections represent unit similarity operations. There are several ways how these unit measurements can be combined into a pipeline to produce the final alignment. A simple pipeline chains several terminology- and structure-level matchers or aggregates their similarities into a single value. BLOOMS [47] and WikiMatch [42] are examples of the systems implementing this design choice. In this section, we consider more sophisticated architectures based on iterative or fixed-point computations.

2.6.1 Iterative Algorithms

When aligning highly heterogeneous knowledge repositories, one has to account for situations when an alignment algorithm cannot take an optimal automatic decision. Along with incorporating special rules for dealing with these edge cases, an algorithm can reserve a slot for interacting with a human judge.

Figure 2.4 shows the flow of the PROMPT system mentioned above. To merge two input ontologies, PROMPT computes an initial alignment by comparing titles. The subsequent procedures happen in a loop:

- a user selects an operation suggested by PROPMT. For example, merge classes of two ontologies or copy a category from one ontology into the merged one;
- PROMPT performs the operation and makes needed updates, for instance, after copying a category, all of its children have to be copied;
- an operation can potentially lead to a conflict or inconsistency such as a copied category without its parent category. In these cases, PROMPT generates possible operations to resolve the conflict (e.g., copying the parent category too) and pushes them into the operations queue.

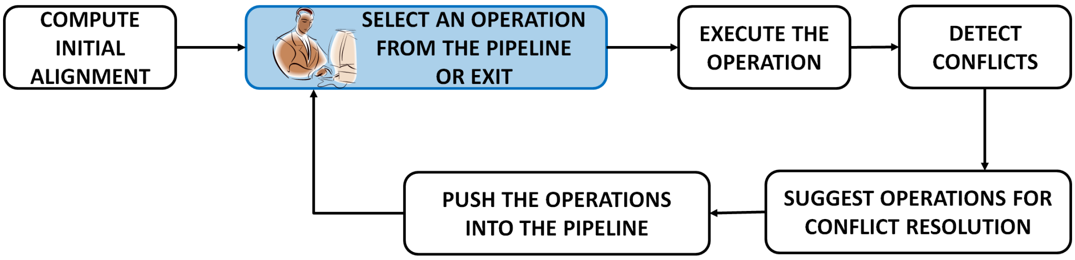

COMPUTE INITIAL ALIGNMENT
SELECT AN OPERATION FROM THE PIPELINE OR EXIT
EXECUTE THE OPERATION
DETECT CONFLICTS
SUGGEST OPERATIONS FOR CONFLICT RESOLUTION
PUSH THE OPERATIONS INTO THE PIPELINE

Figure 2.4: **Iterative alignment computation implemented in PROMPT. The operation in the blue box is performed by the user. The white boxes denote automatic procedures.**

This model actively involves a user into the merging cycle. However, tools of this kind have to expose the provenance of the inconsistencies or conflicts in a very clear way such that a user can take an appropriate decision about its resolution.

2.6.2 Fixed Point Algorithms

In contrast to the iterative approach, fixed point algorithms have a numerical criteria for terminating alignment updates.

PARIS [96] is a probabilistic framework for aligning relations, instances and schemas. The main idea of PARIS is that knowing instance equivalences contributes to the confidence about similarity of relations and, in return, equivalent relations help to detect equivalent instances. An intuitive way of computing probability that two relations r and r' are equivalent, is to compute the portion of instance pairs shared between these relations:

$$P(r \subseteq r') = \frac{|\{i, j : r(i, j) \wedge r'(i, j)\}|}{|\{i, j : r(i, j)\}|} \tag{2.11}$$

Assuming that equivalent relations between two ontologies are known, the probability of instance equivalences can be written in the following form:

$$P(i \equiv i') = 1 - \prod_{r(i,j), r(i',j')} (1 - fun^{-1}(r) \cdot P(j \equiv j')) \tag{2.12}$$

where $fun^{-1}(r)$ is global inverse functionality of relation r. A functionality of a relation is the average number of objects j per subject i:

$$fun(r) = \frac{|\{i : \exists j : r(i, j)\}|}{|\{i, j : r(i, j)\}|} \tag{2.13}$$

For example, on average a person in an ontology has one nationality. The inverse functionality is the functionality of the reversed relation: $fun^{-1}(r) = fun(r^{-1})$.

Figure 2.5 illustrates the loop of computations performed by PARIS. Upfront, the similarities between instance and relations titles are computed using simple literal functions and for each pair of relations r and r' the probability $P(r \subseteq r')$ is set to θ. PARIS updates probabilities of equivalences of instances and relations in turns. At each iteration, the assignments with maximal weights are fixed and used in the next step. Once these assignments stop changing, the algorithm is said to have reached its fixed point and stops. To this end, the

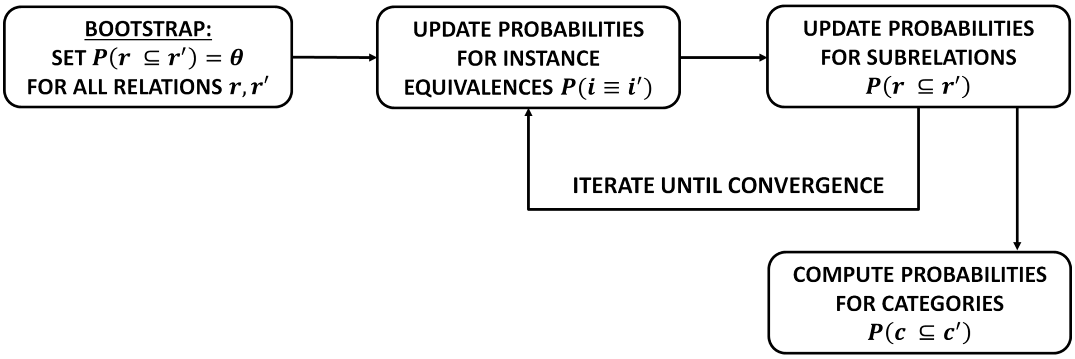

BOOTSTRAP:
SET $P(r \subseteq r') = \theta$
FOR ALL RELATIONS r, r'
UPDATE PROBABILITIES FOR INSTANCE EQUIVALENCES $P(i \equiv i')$
UPDATE PROBABILITIES FOR SUBRELATIONS $P(r \subseteq r')$
ITERATE UNTIL CONVERGENCE
COMPUTE PROBABILITIES FOR CATEGORIES $P(c \subseteq c')$

Figure 2.5: **Fixed point alignment computation implemented in PARIS.**

probabilities for equivalences of categories are computed based on the number of shared instances.

2.7 Conclusion

This chapter gave an overview of different kinds of knowledge repositories: taxonomies, link graphs and ontologies. Then, we defined the task of alignment construction for a pair of knowledge taxonomies. When an alignment is constructed, a knowledge repository curation can take place, i.e., learning missing associations from aligned KR. We gave an outline of techniques to compute alignments between KR objects using terminological and structural information. Lastly, we reviewed two models which combine the unit similarity computations into a pipeline to produce the final alignment.

Chapter 3

ACROSS: Reducing Terminological Heterogeneity

One of the challenges in enabling an interoperability of independently created knowledge repositories is bridging the terminological gap. This means, to determine the identity of the categories and instances contained in a knowledge repository, language varieties and possible multilingualism should be taken into account. If the terminological gap is too large, solutions such as converting category or instance titles to a canonical language via a translation tool or mapping onto a synonymy source such as WordNet are of limited advantage. For illustration, consider a book category with ancient Greek writings *Ancient & Classical*. Its possible counterparts in another catalog are *Classical Greek Poetry* and *Classical Greek speeches*. Detecting relevant counterparts and ranking them is a challenging task. In this chapter, we introduce the ACROSS[1] system, our contribution to supporting alignment of knowledge repositories with high terminological diversity.

3.1 Motivation

The terminology used in a knowledge repository is shaped by its unique culture, which has manifold features. One of them is *language*. The importance of providing the data and products for users globally has been recognized by organizations in the public domain and e-commerce:

> 43% of Europeans never purchase online products and services in languages other than their own[2].

[1] Short for ACcuRate alignment of multi-cultural taxOnomy SystemS

[2] `http://www.lr-coordination.eu/multilingual-europe`, accessed on 28.02.2017

To address this, online marketplaces concentrate efforts on localizing their user interfaces. Initiatives such as the European Language Resource Coordination[3] launched by the European Commission have their objective in developing methods for automated translation to enable public services across borders of the member states. Both use cases refer to offering a unified access to the same data collection through many languages. We are primarily interested in interlinking objects originating from different knowledge repositories.

Another cultural facet of a knowledge repository is the *purpose of its creation and its market orientation.* Organization of product catalogues is motivated by retail success. Among other goals, online marketplaces shape their classifications to boost sales of unsought products (i.e., the products the buyers were not aware of) or to increase the visibility of related products to enable their comparison. In contrast, the objective of library classification systems such as Dewey Decimal Classification (DDC) is to organize locations within a library according to subjects. This leads to a totally different organization of, for example, books with ancient Greek poetry:

> Amazon.com: *Dramas & Plays/Ancient & Classical*
>
> DDC: *Literature & Rhetoric/Hellenic Literatures, Classical Greek/Classical Greek Poetry*

Existing methods in ontology alignment and data integration assume high affinity between the two schemas to align. When aligning knowledge repositories mined Web-wide, high affinity can not be guaranteed.

As a concrete usage scenario, consider a book lover who is interested in finding out which books like-minded people are associating with his favorite topic in a different language of Amazon's online shop or on a social tagging site such as Shelfari. Ideally, a user gets a handful of most relevant categories in a counterpart knowledge repository to which he can navigate. An alignment between categorization systems of online shops and social tagging sites allows for a seamless transition between them.

The considerations outlined in this section motivate the demand in an alignment method which:

1. produces a match between two categories if they are referred to using different vocabulary or language, and

2. outputs a concise and accurate list of relevant counterparts.

[3] `http://www.lr-coordination.eu/` accessed on 27.06.2017

To address these challenges, we build a two staged alignment pipeline within ACROSS system.

The first step – *semantic enrichment* – is the core of ACROSS. Via mapping onto an intermediate taxonomy (Wikipedia), categories of both knowledge repositories are annotated with semantic labels. By computing similarities over the semantic labels, pairwise correspondences between categories of two knowledge repositories can be found.

The second step – *constraint-aware reasoning* – ensures linking to the most semantically related nodes while respecting two types of constraints. A hierarchy-preserving rule disallows that a descendant of a node c in one taxonomy is mapped to an ancestor of c's counterpart in the other taxonomy. Another rule ensures the coherence of the counterpart candidate sets by filtering out non-correlating candidates.

3.2 Contribution

ACROSS system addresses the outlined problem of alignment construction over knowledge repositories with high terminological heterogeneity. Reducing the terminological gap is a highly relevant task for applications which operate with many independently created knowledge repositories. The ability to jointly explore multiple KRs of this kind places the enormous asset of cultural diversity at the users' disposal.

For mapping categories between different taxonomies, ACROSS harnesses instance-level features as well as distant supervision from an intermediate source like multiple Wikipedia editions. ACROSS also includes a reasoning step to arrive at high-quality alignment, using integer linear programming.

In summary, the contributions presented in this chapter are:

- *multi-cultural alignment*: we define and model the alignment problem for multi-cultural knowledge taxonomies;

- *background knowledge*: we utilize a taxonomy mediation source for category assignment of culture-independent semantic labels;

- *alignment as integer optimization problem*: we develop an algorithm for computing alignments based on the semantic labels, using integer optimization;

- *user involvement*: we study different seeding strategies for bringing the run-times down for exact reasoning with two types of constraints, without sacrificing the quality of the alignment;
- *comprehensive experimental study*: using user assessments for alignments between a variety of KB pairs we:
    - analyze linkings produced by ACROSS with respect to concepts with high and low spelling differences. We demonstrate that ACROSS is able to cover more cases, where relying on syntactic similarity or translation fails;
    - perform sensitivity study of linking with respect to the taxonomic levels. ACROSS outperforms the baseline solutions, producing linkings for categories on all taxonomic levels;
    - demonstrate that the proposed seeding strategies drastically reduce the run times of the reasoning step when dealing with complex taxonomies.

3.3 Computational Model

Definition 2.2 creates the basis for the computational model used in ACROSS. Given two independently created taxonomies T_1 and T_2, ACROSS considers:

1. the sets of categories $C(T_1)$, $C(T_2)$;
2. the sets of instances $I(T_1)$, $I(T_2)$;
3. *i-c* edges for connecting instances to their categories; and
4. *c-c* edges for expressing parent-child relation between two categories.

As we outlined in the introduction to this chapter, the heart of ACROSS is in bridging the terminological gap via mapping categories of T_1 and T_2 onto an intermediate taxonomy. In the intermediate taxonomy $\mathcal{I}$ we regard only the sets of its categories and instances – $C(\mathcal{I})$ and $I(\mathcal{I})$, and dismiss all the edges.

Problem Definition 3.1. Let T_1 and T_2 be two taxonomies with high terminological heterogeneity. Our goal is to find an alignment $\mathcal{A} = \{\langle c, k, conf \rangle : conf \in \mathbb{R}\}$ with the help of $C(\mathcal{I})$ and $I(\mathcal{I})$. More specifically, the goal is to compute, for each category c of T_1 a ranked list of most suitable counterparts $k_1, k_2, \ldots$ in T_2. An analogous list is constructed for each $k \in C(T_2)$ as well.

3.4 Overview of ACROSS

Figure 3.1 illustrates the alignment procedure implemented in ACROSS.

We harness Wikipedia as an intermediate knowledge taxonomy, as different Wikipedia editions offer pages from a variety of languages. Throughout this chapter, we collectively refer to articles and categories in Wikipedia as pages.

We can associate a category node c from a given taxonomy T with a set of Wikipedia pages, using simple mapping heuristics onto Wikipedia, either based on the instances of c or based on the surface name of c. For this mapping, we choose the Wikipedia edition that corresponds to T's language, as illustrated in the example in Figure 3.1. *Historical Novels* and *Historischer Roman* are two labels obtained from Wikipedia. We canonicalize the labels towards one of the Wikipedia editions by following the inter-language links.

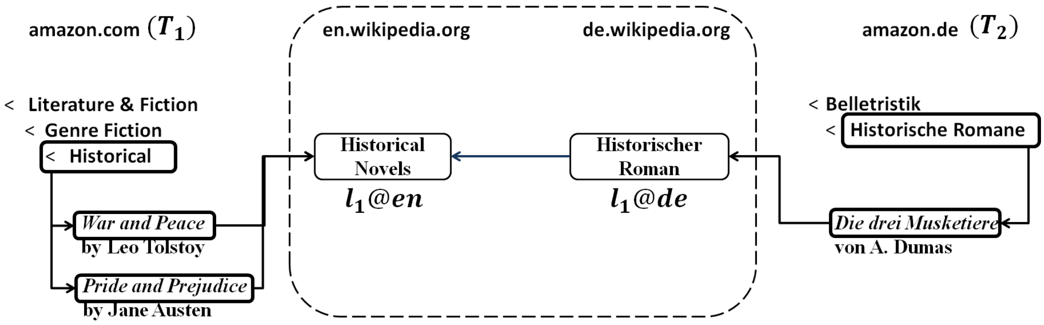


Figure 3.1: **Example alignment of categories** *Historical* **from amazon.com and** *Historische Romane* **(Historical Novels) from amazon.de. Wikipedia serves as a mediator for obtaining labels.**

Note, that instead of Wikipedia any Wikipedia-like source can be deployed. To this end, by comparing the sets of associated Wikipedia pages we compute pairwise similarities between categories of two taxonomies and produce an alignment $\mathcal{A}$.

In summary, given two taxonomies T_1 and T_2, we compute an alignment and resolve terminological heterogeneity in three major steps:

1. Compute semantic labels for all categories c and k of T_1 and T_2, respectively, via mappings to an intermediate Wikipedia edition by finding relevant Wikipedia pages for

   a. the titles of c and k

b. the instances of c and k

The titles of the relevant Wikipedia pages are considered as semantic labels. To perform matching onto Wikipedia pages, we rely on the Wikipedia search API. Contrary to using lexical rule-based matching strategies, we do not depend on the language in which the matching is carried out. Based on the overlap of the semantic labels, the instances-based and name-based alignments are produced.

2. Generate candidate mappings between T_1 and T_2 by combining instances-based and name-based mappings.

3. Consider additional constraints on the alignments and use combinatorial optimization methods to identify good alignments among the candidate ones.

Steps 1 and 2 can be viewed already as complete albeit very basic alignment algorithm. Step 3 performs constraint-aware reasoning to produce concise and accurate alignment suitable for human consumption. We discuss this step further in Section 3.7.

3.5 Semantic Labels (Step 1)

3.5.1 Name-based Semantic Labels (Step 1.a)

The name-based rule finds relevant Wikipedia pages for category c of taxonomy T using the title of c.

Definition 3.1. Let L_c be a set of title-based semantic labels for category $c \in C(T_1)$ and L_k the analogous set for category $k \in C(T_2)$. Then *name-based similarity n-sim*(c,k) between c and k is defined as Jaccard similarity between L_c and L_k:

$$n\text{-}sim(c,k) = \frac{|L_c \cap L_k|}{|L_c \cup L_k|}$$

In this scenario, search results and their socially curated inter-language links are used as a “smart translation” of the Wikipedia community. If two category titles are not exact translations of each other, the alignment between them still can be restored. As an example, a category from medical department of amazon.de *Blutzuckermessgeräte* (en. Glucometers) can be matched with the

English target *Blood Glucose Monitors* without being a literal translation and without involving expensive synonym resolution procedure.

3.5.2 Instance-based Semantic Labels (Step 1.b)

In contrast to the name-based procedure, we pose instance names from a taxonomy to the Wikipedia search API and retrieve a list of relevant pages per instance. Wikipedia search results serve as *semantic labels* for the instances and, transitively, for the categories in T_1 and T_2. In the case of two taxonomies T_1 and T_2 originated in different languages, search results are canonicalized to one of the both languages. We achieve this by following the inter-language links in Wikipedia.

Contrary to the name-based rule, the same semantic label can be assigned to a category through many instances. In Figure 3.1, two instances of the same category return *Historical Novels* in the search. A natural way of modelling this situation is expressing each category c in the taxonomy T as a frequency vector over the set of semantic labels. A frequency vector captures the weight of a semantic label in a category, as well as its specificity – distribution over all categories in the source. This is similar to the $TFIDF$ measure for terms in a document collection.

Let $V_c = \langle v_{c,1}, v_{c,2} ... \rangle$ be the frequency vector of semantic labels for category c. Each component $v_{c,l}$, describing label l, is computed as:

$$v_{c,l} = lf(l,c) \cdot icf(l,T) \tag{3.1}$$

with $lf(l,c)$ being the label frequency in category c and $icf(l,T)$ being the inverse category frequency in source T.

$$icf(l,T) = \log \frac{|C(T)|}{C'} \tag{3.2}$$

where C' is the number of categories containing l.

Due to following inter-language links, categories from T_1 and T_2 are mapped to the same space of semantic labels.

Definition 3.2. The *instance-based similarity* of two categories $c \in C(T_1)$ and $k \in C(T_2)$ is defined as cosine similarity over their frequency vectors of semantic labels:

$$i\text{-}sim(c,k) = \frac{\sum_{l=1}^{n} v_{c,l} \cdot v_{k,l}}{\sqrt{\sum_{l=1}^{n} v_{c,l}^2} \cdot \sqrt{\sum_{l=1}^{n} v_{k,l}^2}} \tag{3.3}$$

where n is the total number of semantic labels.

In contrasts to the Definition 3.1, semantic labels contribute to categories with different weights. To account for the weights we have chosen cosine similarity as one of the standard approaches.

Using Wikipedia search accounts for linguistic complexity, niche- and market-specific instances. Drug names are a good illustration, as they are usually not shared across countries. Consider two categories - *Pain Relievers* from a U.S.-based retailer and *Schmerzmittel* (en.: Pain Relievers) from a Germany-based one. *Aleve* is a product in *Pain Relievers* and *Dolormin* is a product in *Schmerzmittel.* In this representation, both categories contain disjoint set of products. Through mapping to Wikipedia pages, both drug names are lifted to the semantic label *Naproxen.* This lifting allows “crossing” the market borders and making a transition between categories *Pain Relievers* and *Schmerzmittel.*

3.6 Candidate Alignments (Step 2)

The second step merges mapping produced by instance- and name-based rules. Alignment weight $w(c,k)$ between categories $c \in C(T_1)$ and $k \in C(T_2)$ is a linear combination of two weights:

$$w(c,k) = \alpha \cdot i\text{-}sim(c,k) + (1-\alpha) \cdot n\text{-}sim(c,k) \tag{3.4}$$

For a source $c \in C(T_1)$, the found candidate targets $k_1, k_2...$ are ranked according to their weights. Parameter α controls which of the two semantification rules is more emphasized. In our experiments, we used $\alpha = 0.5$.

Combining the two rules induces benefits in at least two aspects. First, we reduce the problem of *sparsity.* This occurs, when a category has a long or rare title and the name-based rule fails to generate a mapping, the instance-based mapping still produces an alignment. Second, we apply community knowledge in order to resolve textual *ambiguities.* We achieve this by incorporating the weights coming from the instances-based mapping. Instances serve as a context for ranking categories with ambiguous names. For example, both book categories *Fiction by Country/Germany* and *Travel/Germany* have the same category name, but can be clearly disambiguated while fetching semantic labels from their instances.

Definition 3.3. Thus, we compile a set of candidate alignments $\mathcal{A}'$ by thresholding with a user-defined value θ:

$$\mathcal{A}' = \{\langle c, k, w(c,k) \rangle : w(c,k) \geq \theta\} \tag{3.5}$$

3.7 Constraint-Aware Reasoning (Step 3)

The basic alignment described in the previous section maps each source category $c \in C(T_1)$ to a set of candidate targets $k_1, k_2 ... \in C(T_2)$ in isolation. This can lead to redundant or erroneous alignments – e.g. mapping to both, a parent and a child category, or mapping to two uncorrelated categories. The methods introduced in this section are aimed at joined alignment between a pair of taxonomies. We formulate the reasoning task in terms of integer linear programming (ILP).

The variables of the ILP model are created as follows. For each triple $\langle c, k, w(c,k) \rangle \in \mathcal{A}'$ we create a binary variable $A_{c,k}$. $A_{c,k}$ is set to 1 if categories c and k are aligned in the current solution. Otherwise, it is 0.

3.7.1 Objective Function

The primary goal is to find an alignment with the maximal weight. Linking between a pair of categories c and k, from T_1 and T_2 respectively, is weighted as in Formula 3.4. When considering all candidate linkings between T_1 and T_2, the objective is:

$$\max \sum_{\substack{c \in C(T_1), \\ k \in C(T_2)}} w(c,k) \cdot A_{c,k} \tag{3.6}$$

It is obvious, that by setting all $A_{c,k} = 1$, the function reaches its maximal weight and we get the original alignment $\mathcal{A}'$. This, however, can lead to inconsistent alignments. In the next subsection we describe two types of inconsistencies and introduce constraints to counter them.

3.7.2 Constraints

Child Constraint (PCH)

Taxonomies organize their categories in hierarchies. When mapping different source categories to a target taxonomy, we could arrive at a situation where a parent-child relationship in the source taxonomy is reversed in the mapping to the target taxonomy. Figure 3.2(a) shows an example. We view such a situation as a violation of a parent-child constraint. We consider two cases:

a). A source category c is linked to targets k and k', and k is a (transitive) parent of k'. For example, *Literature & Fiction* from amazon.com might be linked both to *Belletristik* and *Belletristik/Historische Romane*. Dropping the latter target category makes the candidate list more concise.

b). There is a pair of crossing links – a parent-child pair from the source taxonomy is linked to a child-parent pair in the target taxonomy. In this case, only one of the two linkages should be kept. In the example of Figure 3.2(a), aligning the pair of categories (*Historical*, *Historische Romane*) should exclude the noisy pair (*Short Stories*, *Belletristik*) from a feasible solution.

We introduce a set of linear constraints in order to exclude the alignments violating the hierarchy relation. Expression 3.7 blocks linking category c both to k and k's parent. Thus, it tackles the violation of type a.

$$\begin{aligned} &\forall c \in C(T_1), k, k' \in C(T_2): \\ &k \text{ is more general than } k' \\ &A_{c,k} + A_{c,k'} \leq 1 \end{aligned} \tag{3.7}$$

The analogous constraint is added for a category $k \in C(T_2)$ and a pair of categories $c, c' \in C(T_1)$, where c is more general than c'.

In order to resolve the violation of type b, at most one linking from a pair of crossing links might enter a feasible solution.

$$\begin{aligned} &\forall c, c' \in C(T_1), k, k' \in C(T_2): \\ &c \text{ is more general than } c' \text{ and} \\ &k' \text{ is more general than } k \\ &A_{c,k} + A_{c',k'} \leq 1 \end{aligned} \tag{3.8}$$

Anti-Correlation Hard Constraint (ACH)

This set of constraints addresses another desirable property of taxonomy alignments. When mapping a source category c to multiple target categories $k_1, k_2 \ldots$, we expect the target categories to be semantically coherent. Figure 3.2(b) illustrates a situation where this is violated. Candidate target *Computer & Internet*, which is obviously a wrong match, is negatively correlated with the other two candidate targets. Dropping it makes the candidate list more coherent.

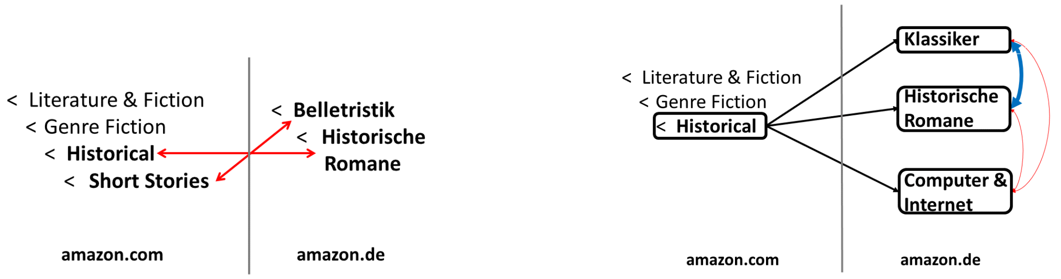
We formalize this intuition by computing the instance-based correlation between candidate targets. When two targets are negatively correlated, only one of them should be kept. This is specified by the following constraints:

$$\begin{aligned} A_{c,k} + A_{c,k'} &\leq 1 \text{ if } corr(k,k') \leq 0 \\ A_{c,k} + A_{c',k} &\leq 1 \text{ if } corr(c,c') \leq 0 \end{aligned} \tag{3.9}$$

where $corr(x,y)$ is the Pearson's correlation coefficient between the instance vectors of the categories x and y:

$$corr(x,y) = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{3.10}$$

x_i expresses the number of occurrences of instance i in the category x to capture multiple occurrences of an instance in a category. Entries of y have analogous meaning.



(a) **Parent-child constraint violation.**

(b) **Anti-correlation constraint violation. Counterpart** *Computer & Internet* **correlates negatively with the rest of the linkings (***Klassiker* **and** *Historische Romane***).**

Figure 3.2: **Examples of constraint violations.**

Anti-Correlation Soft Constraint (ACS)

Forcing all candidate targets to be positively correlated may be too aggressive. Instead, we can relax the anti-correlation constraint and define a “soft” variant of it via a penalty or reward term in the objective function of the combinatorial optimization.

For a source category c and candidate targets k_1, k_2... the reward is the pairwise correlation between all target categories. We denote this by $corr_{T_2}$:

$$corr_{T_2/T_1} = \sum_{c \in C(T_1)} \sum_{k \in C(T_2)} \sum_{k' \in C(T_2)} corr(k, k') \cdot A_{c,k} \cdot A_{c,k'} \tag{3.11}$$

In other words, $corr_{T_2/T_1}$ expresses the degree of coherence within the taxonomy T_2 when matching the classes of T_1 to the classes of T_2.

Analogously, we define the reward for pairwise correlation of the source categories that would be aligned with the same target. We denote this as $corr_{T_1/T_2}$. Note that negative correlations between category pairs in either the targets in T_2 or the sources in T_1 automatically reduces the value of the sum and thus results in a penalty.

Now, we extend the objective function, beyond merely maximizing the alignment weight, by maximizing the sum of the alignment weight and the two reward terms. The objective function of this model thus becomes:

$$\max[\sum_{\substack{c \in C(T_1), \\ k \in C(T_2)}} w(c, k) \cdot A_{c,k} + corr_{T_1/T_2} + corr_{T_2/T_1}] \tag{3.12}$$

Note that the reward terms contain a product of decision variables. Since A variables are binary, one can easily convert this model into a linear model with linear constraints by introducing a new binary variable for each pair $(A_{c,k}, A_{c,k'})$. It increases the dimensionality of the model, but makes it more expressive. Most of the state-of-the-art solvers like Gurobi are capable to deal with quadratic constraints and/or objective terms and aim to tighten the model formulation by, for example, presolving it and applying cutting planes algorithms[4].

Definition 3.4. The cleaned alignment $\mathcal{A}$ is composed of all the variables $A_{c,k}$ set to 1 after the optimization:

$$\mathcal{A} = \{\langle c, k, w(c, k)\rangle \in \mathcal{A}' : A_{c,k} = 1\} \tag{3.13}$$

[4] `http://www.gurobi.com/resources/getting-started/mip-basics` accessed on 27.06.2017

3.8 Seeding Strategies

All the proposed methods are instances of an integer linear programming, which is known to be NP-hard in general. One way of dealing with large optimization instances is to solve a relaxation of the model. However, if we target the exact solution of the original problem, other approaches have to be studied. Consider an example of a model with two constraints:

$$A_{c,k} + A_{c,k'} \leq 1 \tag{3.14}$$

$$A_{c,k} + A_{l,m} \leq 1 \tag{3.15}$$

The variables in the model are closely coupled by being combined in mutual exclusion constraints. Fixing a variable to value 1 propagates the computation of other variables in the model in a cascading manner. We propose to incorporate a small number of truth linkings into the reasoning model, guiding the solver towards the optimal solution.

Definition 3.5. We define seeds as categories of the source or target taxonomies, for which the perfectly matching counterparts are provided by a human annotator. For example, the pair of categories (*Historische Romane*, *Genre Fiction/Historical*) from the German and the English Amazon match perfectly. Category *Kinder- & Jugendliteratur* of amazon.de is relevant to *Teen & Young Adult* from amazon.com, not matching it perfectly, though.

In previous research the problem of providing a small number of seeds without sacrificing the performance of a classifier has been studied in the scope of semi-supervised learning [17]. In the context of the label propagation framework, seeds are nodes for which correct labels are provided. Lin and Cohen [59] study the impact of selecting seeds based on network properties. The observation is, that “central” (or authoritative) nodes likely spread their influence in the network, so that annotating them will significantly improve the quality of a classifier.

In our study, we address not only the effectiveness of the seed categories with respect to the reasoning procedure, but also the amount of user involvement needed to find a matching counterpart. Our first observation is that some linkings are easier to detect for a human. On the other hand, seed categories can be scored by their impact in the model and the most influential ones be presented to a human annotator for labelling. Subsections 3.8.1 and 3.8.2 describe these two strategies.

3.8.1 Depth-based Seeding

When browsing through the product categories of online shops, one notices that some labelling decisions can be made instantly. Fig. 3.3 presents categories of two Amazon health departments (Germany- and US-based).

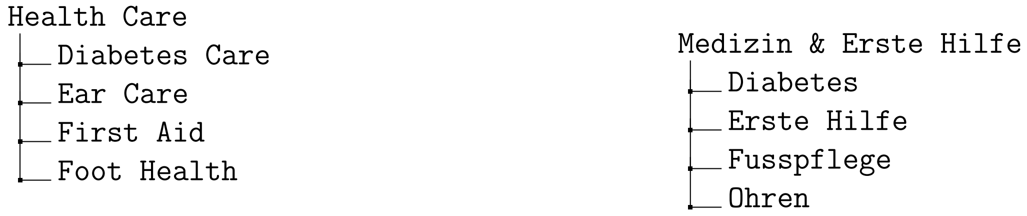


Figure 3.3: **Examples of top level categories in Amazon's health departments (Germany- and US-based).**

The matching categories can be detected by literal translation of category titles, and producing these alignments is not laborious. Generally, we assume that the top-level nodes are easier to annotate than the nodes deep in the taxonomy. This suggests the following strategy.

Definition 3.6. The *depth-based score* of a source category c is its depth in the source taxonomy.

All categories in the source taxonomy are sorted according to their depth in descending order and the top-k nodes are presented to a user for labelling. The ties are broken at random. This seeding rule has its limitations when the top-level categories of both taxonomies are orthogonal. In practice, one might go to the highest level at which a human annotator can make alignment decisions.

3.8.2 Impact-based Seeding

Despite the simplicity of labelling, following the depth-based strategy may have only small impact on run time.

Assume, all seeded categories appear only in one mutual exclusion constraint each. Therefore, by fixing k seed categories, we resolve at most k constraints. However, there might be categories participating in many constraints. The extended influence of these variables make them better seeds with respect to the the optimization model. Detecting the most influential seeds is the idea behind the impact-based seeding strategy. For labelling purposes, the top k categories scored by impact are presented to a human annotator.

Definition 3.7. The *impact-based score* of a source category c is calculated as the number of times the variables related to c participate in constraints.

In Inequality 3.14, both variables connect source category c with targets. For this constraint, $impact(c) = 2$. In Inequality 3.15 variables describe connections for two sources, c and l. Here $impact(c) = 1$ and $impact(l) = 1$. The total impact score for a source variable is summed up over all constraints in the model. By seeding the feedback on category c (e.g., $A_{c,k}$=1) both $A_{c,k'}$ and $A_{l,m}$ get fixed to zero. In contrast, when fixing the ground truth for category l ($A_{l,m} = 1$), only $A_{c,k}$ is resolved to zero.

3.9 Experimental Evaluation

In order to evaluate the alignment quality of our methods, we performed experiments with different taxonomies and human judges for assessment.

3.9.1 Experimental Setup

We experimented with taxonomies covering three domains: health, books and software. Our experiments are based on data retrieved from amazon.com and amazon.de[5]. Amazon.com is the US-centric Web site of Amazon, while amazon.de represents its German "counterpart". Despite being part of the same enterprise, category names and category system are independently maintained and, thus, different.

Source	**Domain**	**# categories**	**# instances**	**Market Bias**
amazon.de (Health)	Health	150	116,000	German
amazon.com (Health)	Health	198	435,000	US
amazon.de (Books)	Books	8,293	962,000	German
amazon.com (Books)	Books	5,846	1,754,000	US
dnb.de	Books	910	1,720,000	German
shelfari.com	Books	12,803	1,173,000	US
amazon.de (Software)	Software	701	125,000	German
amazon.com (Software)	Software	281	100,000	US

Table 3.1: **Properties of the used taxonomies.**

[5] Health domain: "Health Care" and "Medizin & Erste Hilfe"; books domain: "Books" and "Bücher"; software domain: "Software" in both stores

In addition to the aforementioned alignments “within” Amazon, we add two additional data sets for the book domain: a well curated library catalog from the German National Library, dnb.de, based on the Dewey Decimal Classification (DDC). As for contrasting, we incorporate the social tagging community shelfari.com, which is based on a community-created taxonomy. Thus, the taxonomies are very different in nature. First, they have different *curation levels*, ranging from manually curated up to social tagging. Second, they are culture specific based on their different *origin*. Third, they differ in their *sizes* varying from a broad 10,000 categories (shelfari.com) to focused 150 categories (amazon.de, health branch). Table 3.1 summarizes data set properties.

We now describe how the intermediate taxonomies were used. We consider each instance or category title as a query and retrieve relevant Wikipedia pages using its API. We perform both - title and text search. The top k retrieved results become semantic labels (in the experiments we set $k = 5$). From our manual inspection, we observe that setting k larger blows up the set of semantic labels, which are in many cases noisy. When aligning two taxonomies in different languages, labels of the source language are converted to the target language by following inter-language link. If there is no inter-language link for a search result, this Wikipedia page is disregarded.

Three judges participated in manual evaluation of the generated alignments. Each taxonomy pair was evaluated by two of them on a random sample of 100 categories. Alignment output of each method was annotated as *matching* or *wrong*. The annotators were instructed to mark as *matching* all relevant counterparts. I.e., both categories *Classical Hellenic Poetry and Drama* and *Hellenic Literatures* are considered to be matching for *Drama/Greek and Roman*. Cohen’s kappa of the inter-annotator agreement is 0.69, which is considered to be fairly good [52].

3.9.2 Methods

We have the following models under comparison:

1. The **WikiMatch** [42] approach makes a look-up in Wikipedia to align two input taxonomies. For a given category title as input query, it retrieves the results from the Wikipedia search engine. The similarity between two categories is expressed as the Jaccard similarity over the Wikipedia articles returned for each category. Following the inter-language links provided by Wikipedia allows WikiMatch to compare two data sources from different languages.

2. The **S-Match** [36] method reconstructs logical formulas for each category in the taxonomy. For example, category *History/Europe* is converted to the logical formula *History AND Europe*. A correspondence between two categories is found by comparing their logical formulas. We run S-Match with the *Structure Preserving Semantic Matching* option, which respects structural properties such as matching leaves only with leaves and internal nodes only with internal nodes.

3. **Baseline ACROSS** (Section 3.6).

4. **ACROSS** with enabled constraint-aware reasoning (Section 3.7).

5. **ACROSS with seeding** (Section 3.8).

3.9.3 Measures

We introduce the quality measures by which we compare the effectiveness of different alignment methods.

Let S be the set of source categories in the sample set for assessments. For a category $c \in S$, let $Cand(c)$ be the ranked list of target categories that are generated by some method. Categories in $Cand(c)$ are ranked by the alignment weight (see Equation 3.4) in decreasing order.

Since c is linked to a ranked list of target candidates, we consider standard information retrieval measures provided by TREC evaluation script[6].

1. **Mean Reciprocal Rank (MRR).** We are interested in at which position in the ranked list of output categories we see the first match. Let c be a source category and r the rank of a match. Then, the reciprocal rank is $RR(c) = \frac{1}{r}$ If no match exists, then $RR(c) = 0$.

   For a sample of $|S|$ source categories, the MRR value is defined as:

$$MRR = \frac{1}{|S|} \cdot \sum_{c \in S} RR(i) \tag{3.16}$$

2. **Mean Average Precision**(MAP) captures the accumulated precision over all ranked target categories at different recall levels:

$$MAP = \frac{1}{|S|} \sum_{c \in S} \frac{1}{S_c} \sum_{k \in S_i} precision(Cand(i,k)) \tag{3.17}$$

[6] `http://trec.nist.gov/trec_eval` accessed on 27.06.2017

where $|S|$ is the sample size, S_c is the set of correct counterparts for source c and $Cand(c, k)$ is the ranked list of targets for c with cut-off rank k. We report on MAP with cut-off at rank 5.

Note, that if a method did not return any matching candidate, precision value is taken to be zero.

3. **Success@1** measures the portion of sources for which a correct counterpart was produced at rank 1:

$$success@1 = \frac{1}{|S|} \cdot \sum_{c \in S} precision(Cand(c, 1)) \tag{3.18}$$

4. **Utility** is an unnormalized set utility measure, expressing how noisy is the list of retrieved documents. It rewards the method with α points for finding a correct match and penalizes with β points for retrieving an irrelevant counterpart.

   For a source c, the utility of its counterpart list $Cand(c)$ is:

$$\begin{aligned} utility(Cand) = \quad & \alpha \cdot \text{No. of relevant counterparts} - \\ & \beta \cdot \text{No. of non-relevant counterparts} \end{aligned} \tag{3.19}$$

   The final utility score for a method is computed as average utility over S. In our experiments, we set $\alpha = \beta = 1$.

5. **Coverage** expresses the number of source categories which were aligned with at least one matching counterpart.

3.9.4 Setup

All the methods under consideration rely on the similarity between the category titles. This can bias them towards producing *trivial matches* - nearly word-by-word translations. Therefore, we separate annotated examples into two groups - trivial and non-trivial matches and run the evaluation separately. A source category is said to have a trivial match if there is an unambiguous counterpart which can be detected by translation.

Using the Yandex machine translation service[7], we cast all the titles of German taxonomies into English. The tokens of the titles are lemmatized and sorted, such that the matching between *Benjamin Franklin (president)* and *Presidents: Franklin, Benjamin* can be restored. If such a counterpart can not be found, we claim the source being a non-trivial case. Some examples of non-trivial and trivial alignment cases are given in Table 3.5.

Note that, category *Travel Guides/Europe* from the books department belongs to the non-trivial case as well. Although counterparts can be found by simple translation, they are ambiguous: *Cooking by Continent/Europe*, *Religion/Europe* or *Traveling/Europe*. Such categories belong to the non-trivial cases, since there is a need for disambiguation procedure.

3.9.5 Results

The experimental results for different taxonomy pairs are given in Tables 3.2 and 3.3. Results cover the full range of alignments of taxonomies with respect to size, curation level and origin. The plots in Figure 3.4 summarize the percentage of source categories which could be covered by at least one counterpart depending on method and category's depth in the taxonomy.

Across all the experiments, we observe that for non-trivial cases the performance of all methods degrades whereas trivial alignments can be restored by any method with fairly high MAP@5, MRR and success@1 values.

We now discuss our findings on the strengths and weaknesses of each method separately.

WikiMatch outputs high quality alignments in terms of MRR, success@1 and utility for almost all of the use-cases. However, when considering two taxonomies with dissimilar categorization criteria and category titles, WikiMatch does not ensure high coverage for cross-lingual scenarios for non-trivial cases (for example, alignment between amazon.com↔amazon.de on Health). The absence of a reasoning or an alignment repair step leads WikiMatch to incoherent counterparts for category *Religion* in amazon.de (Books) (cf. Table 3.4), where all the counterparts are aligned with the highest weight (1.0). The authors of WikiMatch discuss this limitation as well. WikiMatch also experiences difficulties when aligning categorization schemes with different naming crite-

[7] `https://tech.yandex.com/translate/` accessed on 27.06.2017. Our choice of the translation tool was motivated by the volume of data one can translate using free service. For Yandex it is 10,000,000 characters/month (as compared to 2,000,000 characters/month for Microsoft Translate).

Method	**MAP@5**	**MRR**	**Success@1**	**Utility**	**Relevant Matches**
amazon.de→amazon.com (Health), NON-TRIVIAL CASES					
WikiMatch	0.26	0.40	0.38	-1.6	8
Baseline	0.32	0.37	0.28	-84.00	**62**
ACROSS	0.39	0.44	0.44	**-0.22**	49
+ tree-based seeds	0.38	0.44	0.43	-0.25	46
+ impact-based seeds	0.38	0.45	0.45	-0.26	47
ACROSS SOFT	**0.43**	**0.49**	**0.49**	-0.68	20
amazon.de→amazon.com (Health), TRIVIAL CASES					
WikiMatch	0.78	0.78	0.78	-1	11
Baseline	0.76	0.84	0.78	-93.56	**22**
ACROSS	0.75	0.86	0.86	0.17	20
+ tree-based seeds	0.75	0.84	0.84	0.52	16
+ impact-based seeds	0.80	0.89	0.89	**0.63**	17
ACROSS SOFT	**0.88**	**1.00**	**1.00**	0.61	18
amazon.de→amazon.com (Software), NON-TRIVIAL CASES					
WikiMatch	**0.26**	0.31	0.31	**-1.12**	10
Baseline	**0.26**	0.36	0.27	-35.61	**68**
ACROSS	**0.26**	0.37	0.32	-1.64	44
+ tree-based seeds	0.24	0.33	0.29	-1.64	40
+ impact-based seeds	0.25	0.34	0.30	-1.64	40
ACROSS SOFT	**0.26**	**0.45**	**0.38**	-2.73	35
amazon.de→amazon.com (Software), TRIVIAL CASES					
WikiMatch	0.74	0.79	0.75	**1**	10
Baseline	**0.75**	**0.84**	0.77	-25.5	**17**
ACROSS	0.72	0.80	0.77	0.33	15
+ tree-based seeds	0.68	0.78	0.75	0.31	13
+ impact-based seeds	0.70	0.79	0.76	0.35	14
ACROSS SOFT	0.73	0.83	**0.83**	-0.72	15

Table 3.2: **Experimental results for the domains of Software and Health.**

Method	**MAP@5**	**MRR**	**Success@1**	**Utility**	**Relevant Matches**
amazon.de→amazon.com (Books), NON-TRIVIAL CASES					
WikiMatch	0.14	0.30	0.25	-8.43	24
Baseline	0.11	0.41	0.35	-46.69	**52**
ACROSS	**0.17**	0.36	0.33	**-0.87**	24
+ tree-based seeds	0.10	0.38	0.33	-3.56	26
+ impact-based seeds	0.11	0.41	0.37	-3.13	34
ACROSS SOFT	0.10	**0.49**	**0.47**	-0.94	28
amazon.de→amazon.com (Books), TRIVIAL CASES					
WikiMatch	0.12	0.50	0.46	-1.30	7
Baseline	0.14	0.54	0.46	-88.23	**12**
ACROSS	0.16	0.60	0.60	**0.20**	6
+ tree-based seeds	0.14	0.41	0.33	-7.41	8
+ impact-based seeds	**0.57**	**0.73**	**0.71**	-1.85	6
ACROSS SOFT	0.23	0.72	0.70	-1.60	8
shelfari.com→amazon.com (Books), NON-TRIVIAL CASES					
S-Match	0.47	**0.88**	**0.88**	**0.77**	24
WikiMatch	**0.58**	0.72	0.57	-1.04	**47**
Baseline	0.48	0.61	0.44	-2.38	**47**
ACROSS	0.39	0.61	0.60	0.34	22
+ tree-based seeds	0.45	0.67	0.64	0.57	22
+ impact-based seeds	0.47	0.68	0.66	0.53	23
ACROSS SOFT	0.52	0.73	0.68	-1.25	43
shelfari.com→amazon.com (Books), TRIVIAL CASES					
S-Match	**0.68**	**0.87**	**0.87**	**0.75**	7
WikiMatch	0.62	0.82	0.82	0.63	**38**
Baseline	0.52	0.72	0.71	-1.17	37
ACROSS	0.53	0.75	0.75	0.50	27
+ tree-based seeds	0.55	0.80	0.80	0.60	37
+ impact-based seeds	0.56	0.81	0.81	0.62	26
ACROSS SOFT	0.60	0.82	0.81	-0.40	34
dnb.de→shelfari.com (Books), NON-TRIVIAL CASES					
WikiMatch	0.12	0.38	0.26	-11.23	51
Baseline	**0.29**	0.72	0.64	-1.60	**67**
ACROSS	0.18	0.65	0.62	0.55	49
+ tree-based seeds	0.26	**0.96**	**0.95**	**1.53**	46
+ impact-based seeds	0.24	0.85	0.83	1.10	48
ACROSS SOFT	0.22	0.69	0.68	0.98	50
dnb.de→shelfari.com (Books), TRIVIAL CASES					
WikiMatch	0.19	0.54	0.66	0.41	**17**
Baseline	0.19	0.55	0.40	0.3	**17**
ACROSS	0.13	0.50	0.50	0.25	10
+ tree-based seeds	**0.23**	**1.00**	**1.00**	**1.71**	14
+ impact-based seeds	0.19	0.82	0.82	1.23	14
ACROSS SOFT	0.19	0.70	0.65	1.2	15

Table 3.3: **Experimental results for the Books domain.**

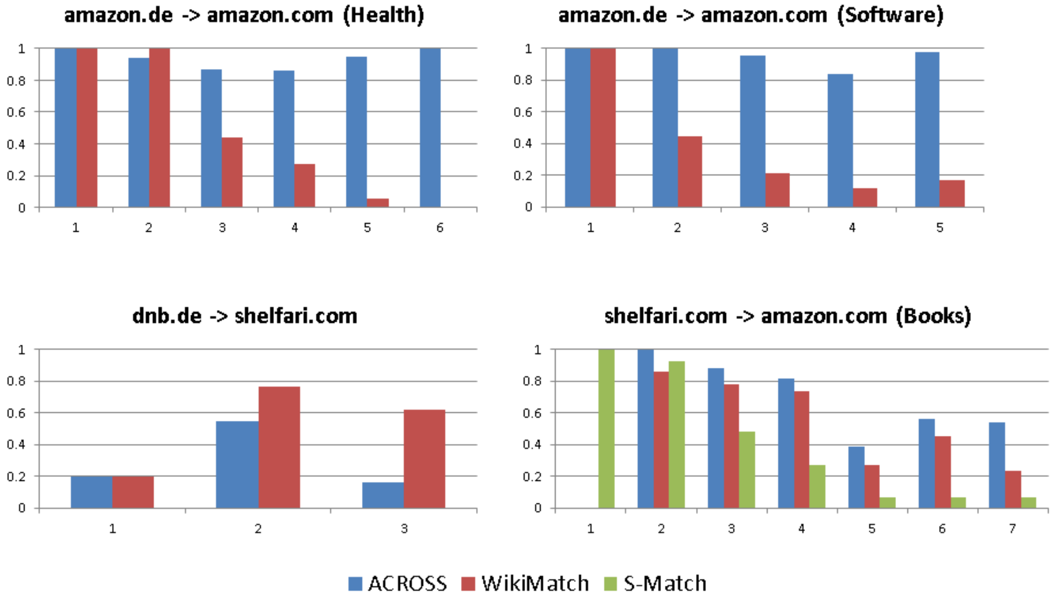


Figure 3.4: **Percentage of source categories per taxonomy level covered by different methods - comparing baseline ACROSS matching with WikiMatch and S-Match .**

ria, e.g. shelfari.com↔dnb.de. The semantic relatedness between categories *Drama/Greek and Roman* and *Classical Hellenic Poetry and Drama* could not be resolved. This leads to lower MAP@5 values for the non-trivial cases.

We run the **S-Match** software on the amazon.com↔shelfari.com use case only, since it is not capable to deal with multi-lingual input taxonomies. S-Match performs well on the sources which have a target with similar tree path and category names along this path, therefore the correct counterpart *History/Europe* is taken and the wrong candidate *Travel/Europe* is eliminated. Slight modifications in wording or tree path decrease recall by filtering out candidates. Dealing with language varieties implies involving additional resources such as WordNet. S-Match ensures non-zero coverage for all levels in the shelfari taxonomy, however only 6% of the leaf categories got matched with a counterpart.

It is worth mentioning that, both, S-Match and WikiMatch, do not consider instances of the categories while constructing an alignment.

Baseline ACROSS. Our baseline solution reaches fairly high MAP and MRR values (up to 0.72 of MRR for dnb.de → shelfari.com for non-trivial cases). Since we rely on the instances for inferring the semantics of a category,

Source Category	**WikiMatch**	**ILP(best configuration)**
Drama/Greek and Roman	–	Classical Hellenic Poetry and Drama Hellenic literatures
Biografien & Erinnerungen/ Religion (en.: Biographies & Memoirs/Religion)	Encyclopedias/Religion Humor & Entertainment/Religion Children's Books/Religions	Biographies & Memoirs/Luther, Martin
Crafts, Hobbies & Home/Scrapbooking	Home and Garden/Scrapbooking	Home and Garden/Scrapbooking
Alternative Medicine/Single Homeopathic Remedies	Homöpathische Einzelwirkstoffe (en.: Homeopathic Individual Active Substances)	Akupunktur (en.: Acupuncture) Alternative Medizin (en.: Alternative Medicine)

Table 3.4: **Anecdotal examples of found alignments.**

Use-case	**Trivial**	**Non-Trivial**
amazon.de↔amazon.com (Health)	Alternative Medizin ↔ Alternative Medicine Erste Hilfe ↔ First Aid	Diabetes/Injektionsspritzen & -kanülen ↔ Insulin Injectors Schlafen & Beruhigung ↔ Sleep & Snoring
amazon.de↔amazon.com (Software)	Betriebssysteme ↔ Operating Systems Sprachen ↔ Languages	Homebanking & Money Management ↔ Budgeting Aktien & Börse ↔ Investment Tools
dnb.de↔shelfari.com (Books)	Sozialwissenschaften ↔ Social Sciences Ethik ↔ Ethics & Morality	Der politische Prozess ↔ Political Theory Bildhauerkunst, Keramik, Metallkunst ↔ Sculpture
shelfari.com↔amazon.com (Books)	Speech Processing ↔ Speech Processing Science & Math ↔ Science	Latin America ↔ Argentina Mountain Biking ↔ Cycling

Table 3.5: **Examples of trivial and non-trivial alignments.**

our basic alignment procedure ensures better coverage. For all the use-cases, all the taxonomic levels could be provided with counterparts, covering more than 90% of the second-level categories in the health, software and books (shelfari.com → amazon.com) experiments. For non-trivial cases in the health domain experiments, the baseline ACROSS outperforms WikiMatch by producing correct alignments for 62 source categories versus 8.

The plots in Figure 3.4 illustrate that the baseline method covers categories on all levels in the taxonomy. In the experiments on Health and Software domain, it produced counterparts for more than 80% of categories on depth 3-6. In the shelfari.com↔amazon.com experiment, both baseline ACROSS and WikiMatch failed to produce the correct linking between roots *All Books* (shelfari.com) and *Books* (amazon.com). The results of related pages returned from Wikipedia search API for queries *All Books* and *Books* are dissimilar, which leads to almost zero Jaccard coefficient. On the instances level, categories contain representative books from all child categories. Therefore, cosine similarity

Setting	**Method**	**Time in sec.**
amazon.de↔amazon.com (Health)	ACROSS	47.85
	+ tree-based seeds	35.40
	+ impact-based seeds	**25.95**
amazon.de↔amazon.com (Software)	ACROSS	74,790.07
	+ tree-based seeds	2,795.72
	+ impact-based seeds	**1,648.67**
dnb.de↔shelfari.com (Books)	ACROSS	0.10
	+ tree-based seeds	**0.05**
	+ impact-based seeds	**0.05**
shelfari.com↔amazon.com (Books)	ACROSS	1,175.34
	+ depth-based seeds	1,071.72
	+ impact-based seeds	**874.93**
amazon.de↔amazon.com (Books)	ACROSS	324,854.24
	+ depth-based seeds	189,624.05
	+ impact-based seeds	**35,810.88**

Table 3.6: **Run time for solving ILP models.**

over instances sets was also below pruning threshold.

ACROSS with enabled constraint-aware reasoning increases the utility(purity) of the counterpart recommendations, reaching 0.55 for the dnb.de→ shelfari.com non-trivial use case. It also provides users with more correct counterparts at rank one, outperforming the baseline by more 16% for amazon.de→ amazon.com (health domain) case over non-trivial instances.

ACROSS with seeding. Table 3.6 illustrates how seeding affects run times in comparison to the ACROSS reasoning without seeds. For all settings, 10 pairs of matching categories (i.e., 10 variables) were provided as seeds with the following total number of variables:

1. amazon.de↔amazon.com (Health) - 11,638
2. amazon.de↔amazon.com (Software) - 13,993
3. amazon.de↔amazon.com (Books) - 49,647
4. shelfari.com↔amazon.com (Books) - 55,170
5. dnb.de↔shelfari.com (Books) - 1,506

Incorporating only a small number of seeds drastically reduces the run times for complex cases, when reasoning has to be run over very noisy data or large taxonomies. For example, in the experiments over software domain, ACROSS had to reason over 22 targets per source category on average, whereas for the shelfari.com → amazon.com only upon 8 targets per source on average. Following the impact-based seed selection strategy had the largest impact on bringing run times down. For the experiments over the Software domain, the run times were reduced by factor 45. In addition, the seeding step slightly improves linking quality, by raising MAP@5 to 0.8 for trivial alignments in the health domain and for utility in almost all experiments (see Tables 3.2 and 3.3).

For all the experiments, ACROSS performed best in terms of MRR and success@1, when the anti-correlation constraint was softened according to the Equation 3.12. When a target candidate got fixed, only a few other targets for the same source may enter the final solution.

The ACROSS SOFT configuration penalizes a solution when non-correlating targets are assigned to a source, rather than aggressively filtering them out. For the health domain, this improves MRR values up to 1.0.

3.10 Related Work

The problem of interlinking multi-cultural taxonomies is adjacent to several directions of past research. This section gives an overview of related work within the four research fields.

Alignment with background knowledge has been harnessed for various matching systems. The WikiMatch system [42] annotates objects (i.e., classes, entities, or properties) of the input ontologies with related Wikipedia pages to produce a match. The annotations are found solely relying on the surface object name. However, WikiMatch does not support any disambiguation procedure to distinguish between equally named objects such as categories *Physics/Reference* and *Psychology/Reference*. ACROSS tackles this problem by additionally analyzing instances and reasoning with rules.

The WeSeE-Match tool [80] performs multilingual ontology alignment based on computing string similarities of the translated titles. In contrast to this strategy, we do not employ any translation tool due to various reasons. Book titles, people names and compound category names limit the effectiveness of translation-based systems. If a concept is aligned to several counterparts, WeSe-Match uses edit distance for reasoning and ranking. ACROSS uses a more sophisticated resolution scheme relying on category correlation and taxonomic structure.

Utilizing either a set of domain-specific taxonomies or a set of ontologies from the LOD cloud is shown to drastically boost matcher's performance especially for lexically heterogeneous input data. Sabou et al. [86] combine a large number of possibly heterogeneous intermediate ontologies to make an alignment decision. They also discuss automated selection of the intermediate ontologies, as opposed to laborious manual procedure. In our settings, ACROSS obtains semantic labels from Wikipedia editions, which are fixed prior to the alignment computation. In line with Sabou's approach, relevant concepts from the intermediate ontologies are retrieved by a search engine. We do not explicitly filter out noisy semantic labels and address this problem by entrusting the Wikipedia search engine and by using weighting schemes like TFIDF. Aleksovski et al. [3] target detecting relations between concepts via relations of their anchors in the intermediate ontologies. To map onto the intermediate ontology, simple lexical heuristics are used. Incorporating several Wikipedia editions when aligning multilingual taxonomies places ACROSS in line with the both approaches, which use an ensemble of intermediate ontologies. Nevertheless, ACROSS has a fundamental difference from the reviewed works. Instead of interlinking plain lists, we are concerned about taxonomy alignment.

Alignment with reasoning. Rule-based alignment cleaning is the core component of many ontology matching applications. ALCOMO [64] is a library providing several alignment debugging procedures. For an inconsistent alignment, ALCOMO automatically detects a minimal repair. Since ALCOMO does not employ a weighting scheme for the alignments, the debugging step explicitly targets removing the minimal set of noisy matchings. ACORSS takes an alternative strategy to filter out wrong alignments taking into consideration alignment weights. LogMap [49] performs the clean-up procedure by removing alignments with minimal weights. We, however, focus on producing the final alignment satisfying the constraints and with the highest possible weight.

On the rapidly evolving Web, the ability to compute an **alignment at scale** becomes highly important. Hu et al. [45] consider the problem of matching two large ontologies. The heart of their approach is in partitioning the input ontologies into blocks based on their structural proximities. The alignments are computed within the blocks and then aggregated. However, this system does not include a reasoning step. Usually, the alignment repair step becomes a bottleneck as a distributed solution of the reasoning algorithm is not trivial to find. ACROSS tackles the scalability problem by fetching a small number of anchor alignments from a human annotator. The basic alignment procedure of ACROSS (i.e., without the reasoning) can be parallelized in a straightforward manner, since matching a pair of categories does not depend on the entire hierarchy. The TACI system [79] targets the catalog integration problem for Web-scale taxonomies containing millions of entities. To arrive at a plausible run time, TACI incorporates a number of heuristics such as considering only top-k candidate categories per product or fixing category assignments for some of the products. This is very similar in spirit to the seeding step of ACROSS. However, TACI targets instance-to-category alignments, which are out of the scope of ACROSS.

Multilingual data and knowledge alignment. Nguyen et al. [74] describe a mapping of Wikipedia infoboxes across different editions. Values of infobox attributes are represented as judiciously constructed feature vectors in the underlying Wikipedia. Following the cross-lingual interwiki links allows two attributes from different languages to be compared. In addition, link-structure similarity, correlation similarity, and infobox types are used to compute alignments between infobox fields. In contrast, our setting focuses on categories, which are disregarded in [74]. Moreover, we address a wide variety of taxonomies beyond Wikipedia. Gracia et al. [40] discuss challenges arising from multi-lingual data in the Linked Open Data cloud. Our work is orthogonal to these issues: we focus on culture-specific category systems, not on RDF triples and entity linkage. Spohr et al [92] describe an approach to multilingual and cross-lingual ontology matching. A set of structural and string similarity features is fed into a support vector machine (SVM) algorithm. We do not use any learning algorithm, respecting structural and textual similarities of aligned categories though.

3.11 Summary

This chapter presented the ACROSS system for reducing terminological heterogeneity when aligning multi-cultural knowledge repositories. The heart of ACROSS is a method which maps all categories jointly and considers constraints to arrive at high-quality mappings, using integer linear programming. To narrow the terminological gap, ACROSS incorporates a search-based semantification procedure to map titles of objects in a KR onto Wikipedia articles of corresponding languages. This procedure makes ACROSS independent of synonym resolution and lexical matching tools. Including a structure-aware reasoner into ACROSS alignment pipeline clearly boosts the quality of the alignments. Additionally, we have studied two approaches to incorporate user feedback in order to limit the run times for our exact reasoning procedure.

Chapter 4

LILIANA and SESAME: Reducing Structural Heterogeneity

So far, we have investigated the alignment of taxonomies reducing their terminological heterogeneity. However, once the overlap of concepts or categories becomes rather small compared to the input repositories, terminology-based approaches like ACROSS reach their limits. As an example you might consider the link graph of general concepts published by the European Statistical Organization (Eurostat) and the link graph of Wikipedia, which in addition includes articles about people, organizations, events etc.

Although general concepts like *Refugee* are likely to be found in both link graphs, the article about *European Migrant Crisis* is missed out in Eurostat. This becomes an important limitation, when there is a demand in going beyond strict equivalences.

In this chapter we present two systems, LILIANA[1] and SESAME[2], for overcoming the problem of small structural overlap. Their core is in utilizing interlinking to the unmatched part of the knowledge repositories to unlock hidden alignments.

This chapter serves illustrative purposes as well. We present two analysis scenarios spanning disconnected knowledge repositories. The LILIANA system allows live linking of Web contents such as news articles with online statistical reports. The SESAME system enables explaining numerical statistics via linking onto Wikipedia pages about related events, people or organizations "causing" or being "affected" by statistical observations.

[1] LIve LInking for online statistic ANAlytics

[2] Statistics Explored via Semantic AlignMEnt

4.1 Motivation

Figures 4.1 and 4.2 clarify what we mean by structural heterogeneity for taxonomies and link graphs.

The European Statistical Organization (Eurostat) serves as a gateway to statistical reports and numerical data. The reports are organized in a topic hierarchy, which is quite flat having at most 4 levels and totalling 40 categories. On the other side, there are wide and deep classification systems such as the type hierarchy of YAGO, which catalogs people, organizations, events etc. In total, YAGO holds more than 350,000 object types. As a collection of governmental statistics, Eurostat is primarily concerned about abstract topics like *Asylum and Migration*. YAGO, in constrast, is highly entity-centric. Take category *Migrant Crises* as an example, which includes only individual events.

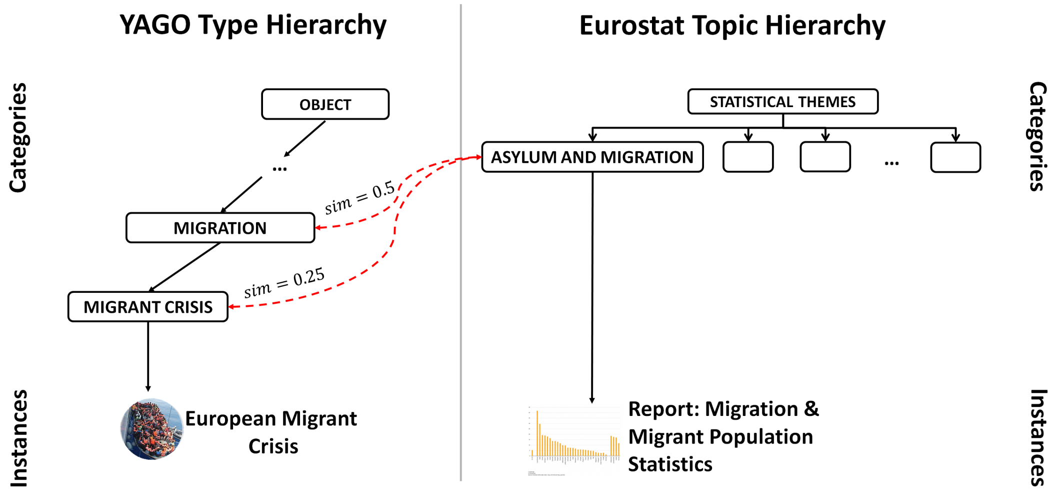


Figure 4.1: **An example of aligning two structurally different taxonomies. The perfectly matching categories are ⟨*Migration, Asylum and Migration*⟩. To extend the alignment, child categories of the matched roots are also aligned but with a discounted confidence.**

Seeking for strict equivalences between categories of two taxonomies is an infeasible solution in these settings. Within our LILIANA system we are looking for a method which enables stepping over the structural gap of two taxonomies. This allows establishing a relevance relation, for example, between categories *Migrant Crises* and *Asylum and Migration*.

Likewise, structural heterogeneity arises between link graphs as illustrated

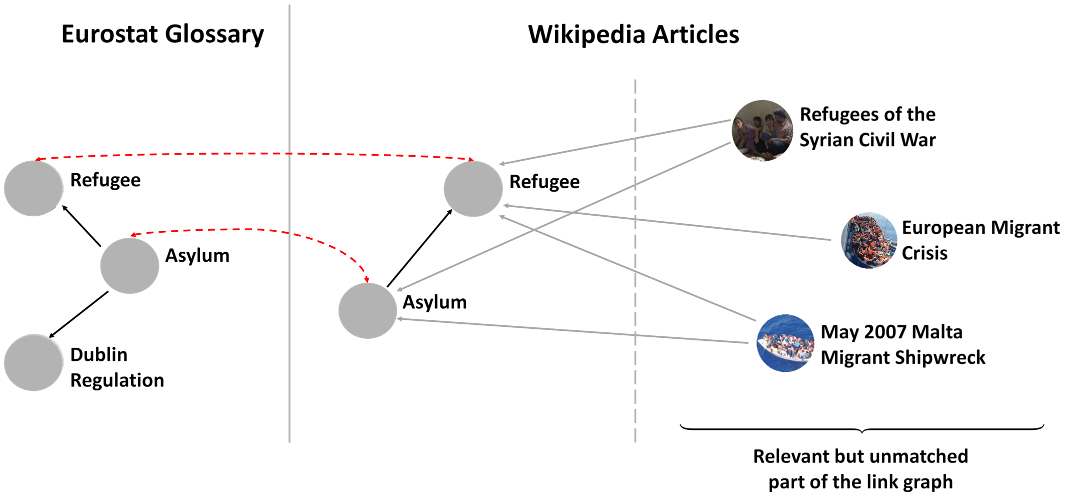

Eurostat Glossary
Wikipedia Articles
Refugee
Asylum
Dublin Regulation
Refugee
Asylum
Refugees of the Syrian Civil War
European Migrant Crisis
May 2007 Malta Migrant Shipwreck
Relevant but unmatched part of the link graph

Figure 4.2: **An example of aligning two structurally different link graphs. The red arrows represent an alignment between perfectly matching nodes of both link graphs.**

in Figure 4.2. Eurostat contains an interlinked collection of glossary terms. These are abstract concepts like *Refugee* and *Asylum*. Wikipedia, in addition to concepts, has a rich collection of entity pages on people or events.

Analogously to the previous use case, a method for enabling a transfer from the instances of one link graph onto related instances of another link graph is sought after. The goal is to include closely related instances such as *European Migrant Crisis* or *Syrians in Germany* when linking from Eurostat's *Refugee*. The second part of this chapter is dedicated to the SESAME system, which targets finding an alignment between structurally heterogeneous link graphs.

In the next two subsections we present two analysis scenarios, where the ability to align structurally heterogeneous knowledge repositories is the core component.

4.1.1 Analysis Scenario 1.

In this scenario, we are interested in providing statistical evidences for Web contents (e.g., news). To explain this by example, assume a user who reads the following excerpt from a news article:

> **Europe migrant crisis: How are countries coping?**[3]
>
> "Europe's migration crisis affects EU member states in different ways - so it is proving difficult to agree on common rules. Germany has more asylum seekers than any other EU country. Its strong economy is a magnet for migrants desperate to start a new life."

In order to understand and to check the validity of the statements made in the news article, contextual information such as online statistics are desirable. The statistical report below supports the claim of the news excerpt:

> **Migration and migrant population statistics**
>
> "In absolute terms, the largest numbers of non-nationals living in the EU Member States on 1 January 2016 were found in Germany (8.7 million persons), the United Kingdom (5.6 million), Italy (5.0 million), Spain (4.4 million) and France (4.4 million)."

Identifying the most suitable statistical document given a specific news article is not a trivial task. Standard approaches that create contextual information by matching keywords or keyphrases are of limited advantage given this settings.

The key idea of the LILIANA system (Section 4.3) is to contextualize input news articles with the YAGO types of mentioned entities. By establishing an alignment between the type hierarchy of YAGO and the Eurostat's topic hierarchy, a set of relevant statistical reports is confined.

4.1.2 Analysis Scenario 2.

In contrast to the claim verification scenario outlined above, there is a need in linking from factual to background knowledge. Indeed, high volumes of factual information become available, e.g. facts contained in the Linked Open Data cloud or highly specialized statistics published by government agencies such as Eurostat[4]. For instance, the reported number of asylum applications over time (cf. Fig. 4.6) shows a significant increase for Germany in the year 2015. However, end users are often left alone with these facts as background information on key concepts (e.g. events, people or organizations) is missing.

As opposed to factual knowledge, crowd-curated Wikipedia contains a wealth of detailed textual descriptions of concepts and entities. For instance, consider

[3] `http://www.bbc.com/news/world-europe-33286393` accessed on 27.06.2017

[4] `http://ec.europa.eu/eurostat/web/regions/data/database` accessed on 27.06.2017

the excerpts from the following Wikipedia articles, where the underlined text denotes internal Wikipedia links:

> **Timeline of the European migrant crisis**
>
> "11–12 November: Valletta Summit on Migration – a summit between European and African leaders was held in Valletta, Malta, to discuss the migrant crisis."
>
> **Horst Seehofer**
>
> "In late 2015, Seehofer and the CSU sharply criticized Chancellor Angela Merkel's refugee policy."
>
> **Wilhelm-Diess-Gymnasium**
>
> "In 2015 accommodations for 200 Syrian refugees from the 2015 [migrant] crisis were established adjacent to the school's gymnasium (athletic facility)."

Obviously, these text snippets associated with an event, a person and an organization are highly relevant to the observed statistical incline in asylum seekers in Germany in the year 2015. The SESAME system (Section 4.4) tackles the problem of detecting correspondences of that kind. The core idea of SESAME is in contextualizing Wikipedia pages with their proximity to the key concepts of the statistical table (such as *Refugee*) within the Wikipedia link graph.

4.2 Contribution

Although LILIANA and SESAME are inspired by different use case scenarios, they have conceptual similarities. Only a small set of alignments between categories or instances can be discovered using terminological alignment tools. To extend the alignment and enable discovering more related objects in a counterpart KR, both systems rely on the structure, either on taxonomy of types and categories or on the link graph over instances.

In this chapter, we summarize our contributions made within two systems.

- LILIANA:
  - *semantic alignment*: we interlink Web (archive) contents with online statistics based on their semantics;

  - *taxonomy-based similarity*: we define similarity between categories of two taxonomies based on the input taxonomic structure;
  - *similarity model* for interlinking Web contents (e.g., news articles) with statistical reports: we incorporate the distance between semantic types detected in the news articles to the categories of the statistical reports. The ranking model can optionally include a text-based similarity.
  - we developed a *graphical user interface and a browser plug-in* for live linking of Web contents to online statistics.

- SESAME:
  - *semantic alignment*: we interlink numerical statistics and Wikipedia articles based on their semantics;
  - *graph-based similarity*: we detect and rank related Wikipedia pages based on their proximity to the domain pages of a statistical table;
  - we propose a *similarity model* for interlinking statistical observations with Wikipedia articles, contextualized with the temporal and spacial dimensions;
  - we develop a *graphical user interface* for jointly exploring numerical statistics and associated Wikipedia articles;
  - we conduct an *experimental evaluation*, showing that SESAME outperforms plain keyword search due to encapsulating semantic similarity into the ranking function.

4.3 Application: LILIANA

In the motivating scenario given in Section 4.1.1, we are interested in an approach for joint browsing of text content such as news articles and statistical reports. As outlined earlier, standard approaches which create contextual information by matching keywords or keyphrases are of limited advantage. Instead, within the LILIANA system we devise a method for semantic retrieval of statistical reports given a snippet of a news article.

To support semantic retrieval, we introduce a three-staged pipeline, which is illustrated in Figure 4.3. Given an input text and a collection of statistical reports classified into a taxonomy of topics, we retrieve a ranked list of statistical reports in the following steps:

1. we lift the input text to the semantic level by detecting entities mentioned in the text and their types registered in the YAGO ontology;

2. we compute an alignment between the taxonomy of the YAGO types and the topic taxonomy of Eurostat; and

3. we cast the types of each entity mention detected in the input text to the corresponding Eurostat topics. These topics confine a set of statistical reports, which we subsequently score.

Step 2 is our main contribution within the LILIANA system. In the next section we formally define the computational model and the alignment task. Sections 4.3.2 to 4.3.4 discuss the outlined algorithmic steps.

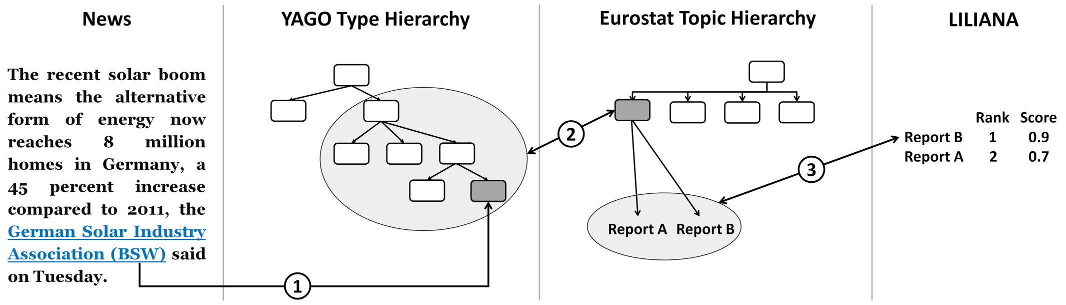

News
YAGO Type Hierarchy
Eurostat Topic Hierarchy
LILIANA
The recent solar boom means the alternative form of energy now reaches 8 million homes in Germany, a 45 percent increase compared to 2011, the German Solar Industry Association (BSW) said on Tuesday.
1
2
3
Report A
Report B
Rank
Score
Report B
1
0.9
Report A
2
0.7

Figure 4.3: **Pipeline of knowledge linking for online statistics.**

4.3.1 Computational Model

Definition 2.2 serves as a basis for the LILIANA's computational model. Consider two taxonomies T_1 and T_2 and the sets of their categories $C(T_1)$ and $C(T_2)$, respectively. There are *c-c* links for connecting a pair of a parent and a child categories.

The key property of the taxonomies is that the number of overlapping categories (nearly duplicates) is very low. This leads to the following problem statement.

Problem Definition 4.1. Let T_1 and T_2 be two taxonomies which have a very small number of overlapping categories. Assume, $\mathcal{A}'$ is an alignment between these overlapping categories: $\mathcal{A}' = \{\langle c_1, c_2, conf\rangle : c_1 \in C(T_1), c_2 \in C(T_2), conf \in \mathbb{R}\}$. Our goal is to find an alignment $\mathcal{A} \supseteq \mathcal{A}'$, which extends the "trivial" alignment $\mathcal{A}'$. That is, for each category $c_1 \in C(T_1)$ which has a trivial match $c_2 \in C(T_2)$, find a list of most relevant counterparts $c_3 \ldots c_n \in C(T_2)$ using the hierarchy of categories in T_2.

In the LILIANA scenario, T_1 is the YAGO type system and T_2 is the topic hierarchy of Eurostat.

4.3.2 Semantic Enrichment (Step 1)

In order to semantically enrich textual Web contents, we lift the plain text to the entity level by detecting named entities and resolving ambiguous names. For this, we employ the AIDA entity disambiguation system [44] that maps mentions of entities onto canonical entities of the YAGO ontology. Each entity in YAGO is associated with a set of types. To give an example, consider the entity *European Central Bank* with its 24 types (*Central Banks* or *Supranational Banks* to mention a few).

4.3.3 Alignment Construction (Step 2)

In order to interpret YAGO types as topics of Eurostat reports, we compute an alignment between YAGO types taxonomy and the topic hierarchy of Eurostat.

Anchor Alignment

As stated in the problem definition, there is a small set of overlapping instances or categories, which is captured in the anchor alignment $\mathcal{A}'$. This set, however, is not given explicitly and has to be constructed first. In LILIANA we rely on token-based methods for assessing similarity between the YAGO types and the Eurostat categories. These methods come from the area of information retrieval and consider a string as a bag of words (tokens).

Let t_1 and t_2 be the titles of a YAGO type c_1 and a Eurostat topic c_2. Similarity between c_1 and c_2 is expressed via similarity between their titles.

The titles undergo tokenization, stemming and removing stop words. Let τ_1 and τ_2 be the resulting bags of words.

Definition 4.1. We define **token-based similarity** between titles t_1 and t_2 as the Jaccard similarity between the corresponding bags of words:

$$sim(t_1, t_2) = \frac{|\tau_1 \cap \tau_2|}{|\tau_1 \cup \tau_2|} \tag{4.1}$$

We map t_1 onto its textually most similar counterpart, requiring that a user-definable threshold θ is exceeded.

Alignment Extension

For knowledge repositories with high structural diversity, the anchor alignment $\mathcal{A}'$ contains only a small portion of categories from $C(T_1)$ and $C(T_2)$. In order to allow for larger discrepancy in structures, the anchor alignment is propagated to the unaligned parts of both taxonomies. Thus, we arrive at an extended alignment $\mathcal{A} \supseteq \mathcal{A}'$.

Let $\langle c_1, c_2, conf \rangle \in \mathcal{A}'$ be a pair of aligned categories. We map category c_1 to the whole sub-hierarchy rooted at c_2 with a discounted confidence based on the distance to c_2. That is, the extended alignment $\mathcal{A}$ includes triples $\langle c_1, c_2', conf' \rangle$, where c_2' is in the sub-hierarchy of c_2 and

$$conf' = conf \cdot \frac{1}{distance(c_2, c_2')} \tag{4.2}$$

Figure 4.1 illustrates this type of extension.

4.3.4 Ranking Statistical Reports (Step 3)

Within LILIANA, we introduce three models for scoring statistical reports given a textual input.

1. **TFIDF.** This is our baseline model, which considers only the textual similarity between the query text q and a statistical report r. They are represented as *TFIDF* vectors over the common dictionary. The relevance of a report to the query text is defined as a cosine similarity between the corresponding vectors $\vec{r}$ and $\vec{q}$:

   $$sim(r, q) = cos(\vec{r}, \vec{q}\,) \tag{4.3}$$

   As we outlined in the motivating scenarios at the beginning of this chapter, this method explicitly prefers documents which have high textual overlap with the query text. This is, however, is rarely the case if two pieces of text have different origin and purpose like news articles and statistical reports.

2. **Voting.** In this setting, we entirely rely on the alignment constructed between the type hierarchy of YAGO and the Eurostat classification. Let

Q_T be the set of YAGO types detected in the query text q, C_r be the set of categories of the statistical report r and $\mathcal{A}$ be an alignment constructed between the YAGO types and the Eurostat topic classification. In the voting method, we directly compare sets Q_T and C_r and prefer statistical reports with the largest number of categories overlapping with the YAGO types derived from the query text:

$$sim(r, q) = |\langle c_r, q_t, conf\rangle \in \mathcal{A} : c_r \in C_r, q_t \in Q_T| \tag{4.4}$$

The voting method, unfortunately, results in many equally ranked documents and a mechanism for breaking the ties has to be devised.

3. **Voting + TFIDF.** This is a hybrid method combining the two previous scoring models. It performs in a two-stage computation. First, we confine the relevant statistical reports based on the voting method and threshold by a user-defined value θ. In the subsequent step, we rank the reports based on their textual proximity to the query text.

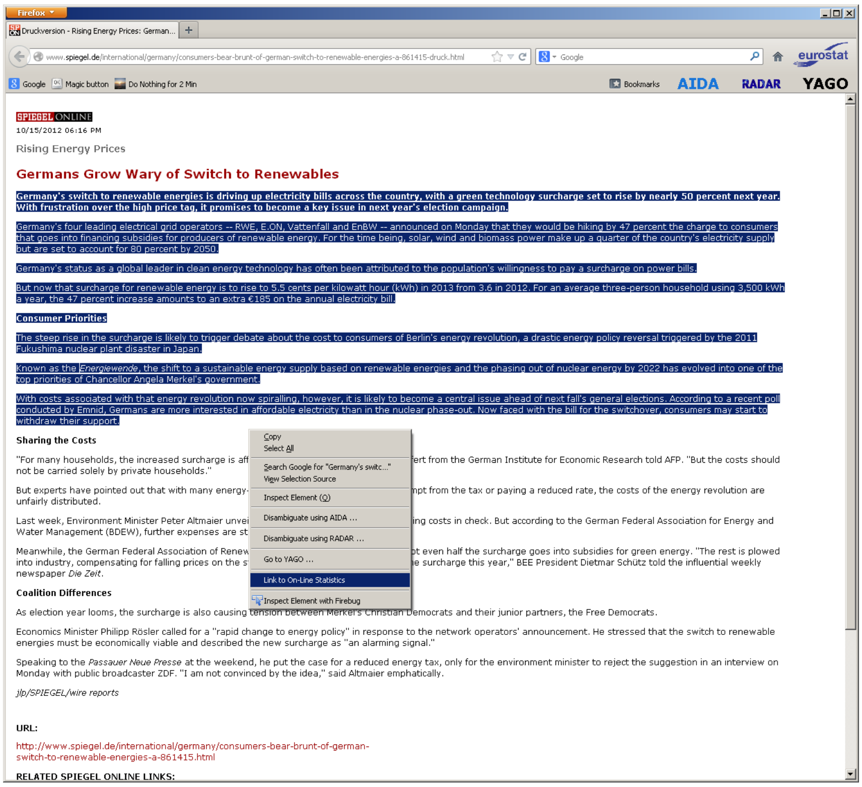
4.3.5 Data

We have chosen the YAGO ontology for semantic enrichment of textual Web contents. It contains more than 10 million entities classified into more than 350.000 types/categories derived from Wikipedia. We obtained the taxonomy, the glossary and the contents of Eurostat by crawling, thus, creating a replica for indexing and alignment. In particular, we have selected the *Statistical Themes* subsection of the hierarchy as it reflects a taxonomic structure used for classification. In total, our Eurostat dataset consists of 40 categories used for classifying almost 2000 statistical articles (English contents only).

4.3.6 Implementation

LILIANA's GUI is Javascript-based. On the server side, LILIANA is written in Java and runs on a Tomcat server. Additionally, we developed a Firefox plug-in, which is written in Javscript. We precompute and store the alignment in a PostgreSQL database.

The interface of the Firefox plug-in is shown in Figure 4.4. It appears as the option *Link to Online Statistics* in the context menu, when some text is selected. By clicking on this option, the user is redirected to the LILIANA GUI and the selected text is used as a query to LILIANA.


Firefox
Druckversion - Rising Energy Prices: German...
www.spiegel.de/international/germany/consumers-bear-brunt-of-german-switch-to-renewable-energies-a-861415-druck.html
Google
eurostat
Google
Magic button
Do Nothing for 2 Min
Bookmarks
AIDA
RADAR
YAGO
SPIEGEL ONLINE
Copy
Select All
Search Google for "Germany's switc..."
View Selection Source
Inspect Element (Q)
Disambiguate using AIDA ...
Disambiguate using RADAR ...
Go to YAGO ...
Link to On-Line Statistics
Inspect Element with Firebug

10/15/2012 06:16 PM

Rising Energy Prices

Germans Grow Wary of Switch to Renewables

Germany's switch to renewable energies is driving up electricity bills across the country, with a green technology surcharge set to rise by nearly 50 percent next year. With frustration over the high price tag, it promises to become a key issue in next year's election campaign.

Germany's four leading electrical grid operators -- RWE, E.ON, Vattenfall and EnBW -- announced on Monday that they would be hiking by 47 percent the charge to consumers that goes into financing subsidies for producers of renewable energy. For the time being, solar, wind and biomass power make up a quarter of the country's electricity supply but are set to account for 80 percent by 2050.

Germany's status as a global leader in clean energy technology has often been attributed to the population's willingness to pay a surcharge on power bills.

But now that surcharge for renewable energy is to rise to 5.5 cents per kilowatt hour (kWh) in 2013 from 3.6 in 2012. For an average three-person household using 3,500 kWh a year, the 47 percent increase amounts to an extra €185 on the annual electricity bill.

Consumer Priorities

The steep rise in the surcharge is likely to trigger debate about the cost to consumers of Berlin's energy revolution, a drastic energy policy reversal triggered by the 2011 Fukushima nuclear plant disaster in Japan.

Known as the *Energiewende*, the shift to a sustainable energy supply based on renewable energies and the phasing out of nuclear energy by 2022 has evolved into one of the top priorities of Chancellor Angela Merkel's government.

With costs associated with that energy revolution now spiralling, however, it is likely to become a central issue ahead of next fall's general elections. According to a recent poll conducted by Emnid, Germans are more interested in affordable electricity than in the nuclear phase-out. Now faced with the bill for the switchover, consumers may start to withdraw their support.

Sharing the Costs

"For many households, the increased surcharge is aff ert from the German Institute for Economic Research told AFP. "But the costs should not be carried solely by private households."

But experts have pointed out that with many energy- mpt from the tax or paying a reduced rate, the costs of the energy revolution are unfairly distributed.

Last week, Environment Minister Peter Altmaier unvei ing costs in check. But according to the German Federal Association for Energy and Water Management (BDEW), further expenses are st

Meanwhile, the German Federal Association of Renew ot even half the surcharge goes into subsidies for green energy. "The rest is plowed into industry, compensating for falling prices on the s e surcharge this year," BEE President Dietmar Schütz told the influential weekly newspaper *Die Zeit*.

Coalition Differences

As election year looms, the surcharge is also causing tension between Merkel's Christian Democrats and their junior partners, the Free Democrats.

Economics Minister Philipp Rösler called for a "rapid change to energy policy" in response to the network operators' announcement. He stressed that the switch to renewable energies must be economically viable and described the new surcharge as "an alarming signal."

Speaking to the *Passauer Neue Presse* at the weekend, he put the case for a reduced energy tax, only for the environment minister to reject the suggestion in an interview on Monday with public broadcaster ZDF. "I am not convinced by the idea," said Altmaier emphatically.

jlp/SPIEGEL/wire reports

URL:

http://www.spiegel.de/international/germany/consumers-bear-brunt-of-german-switch-to-renewable-energies-a-861415.html

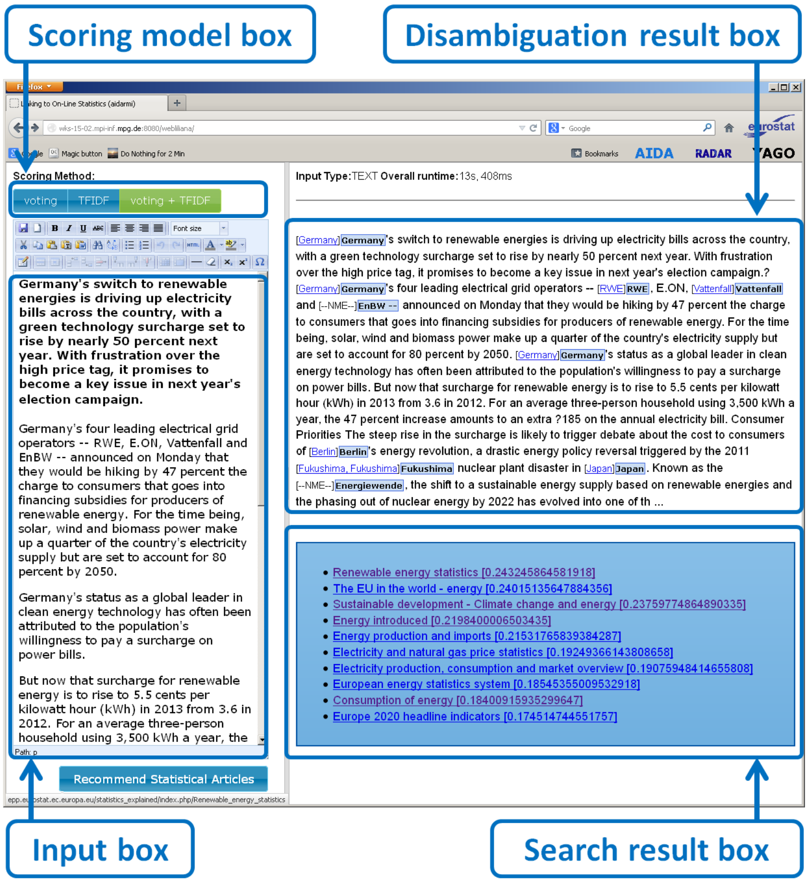
RELATED SPIEGEL ONLINE LINKS:

Figure 4.4: **Overview of the LILIANA Firefox plugin.**

The user interface of LILIANA is shown in Figure 4.5 and comprises the following four key components:

1. **Scoring model box.** Three buttons on the top left panel allow the user to switch between the underlying scoring and retrieval method. As a default setting, LILIANA employs the hybrid approach that combines textual similarity measures with the semantic alignment.

2. **Input box.** The text panel initially contains a copy of the text that has been selected when activating the LILIANA browser plug-in. However, the user may input any text, e.g. by copy-and-paste from arbitrary Web contents, or even HTML tables. By default, the AIDA entity disambiguation system identifies noun phrases that can be interpreted as entity mentions. As this is potentially error-prone, the user can alternatively flag mentions by putting them in double brackets, e.g.: *Harry is the opponent of [[you know who]].*


Scoring model box
Disambiguation result box
Firefox
Linking to On-Line Statistics (aidarmi)
wks-15-02.mpi-inf.mpg.de:8080/webliliana/
Google
eurostat
Google
Magic button
Do Nothing for 2 Min
Bookmarks
AIDA
RADAR
YAGO
voting
TFIDF
voting + TFIDF
Font size
Path: p
Recommend Statistical Articles
epp.eurostat.ec.europa.eu/statistics_explained/index.php/Renewable_energy_statistics
Input box
Search result box

Scoring Method:

Germany's switch to renewable energies is driving up electricity bills across the country, with a green technology surcharge set to rise by nearly 50 percent next year. With frustration over the high price tag, it promises to become a key issue in next year's election campaign.

Germany's four leading electrical grid operators -- RWE, E.ON, Vattenfall and EnBW -- announced on Monday that they would be hiking by 47 percent the charge to consumers that goes into financing subsidies for producers of renewable energy. For the time being, solar, wind and biomass power make up a quarter of the country's electricity supply but are set to account for 80 percent by 2050.

Germany's status as a global leader in clean energy technology has often been attributed to the population's willingness to pay a surcharge on power bills.

But now that surcharge for renewable energy is to rise to 5.5 cents per kilowatt hour (kWh) in 2013 from 3.6 in 2012. For an average three-person household using 3,500 kWh a year, the

Input Type:TEXT **Overall runtime:**13s, 408ms

[Germany]**Germany**'s switch to renewable energies is driving up electricity bills across the country, with a green technology surcharge set to rise by nearly 50 percent next year. With frustration over the high price tag, it promises to become a key issue in next year's election campaign.? [Germany]**Germany**'s four leading electrical grid operators -- [RWE]**RWE**, E.ON, [Vattenfall]**Vattenfall** and [--NME--]**EnBW --** announced on Monday that they would be hiking by 47 percent the charge to consumers that goes into financing subsidies for producers of renewable energy. For the time being, solar, wind and biomass power make up a quarter of the country's electricity supply but are set to account for 80 percent by 2050. [Germany]**Germany**'s status as a global leader in clean energy technology has often been attributed to the population's willingness to pay a surcharge on power bills. But now that surcharge for renewable energy is to rise to 5.5 cents per kilowatt hour (kWh) in 2013 from 3.6 in 2012. For an average three-person household using 3,500 kWh a year, the 47 percent increase amounts to an extra ?185 on the annual electricity bill. Consumer Priorities The steep rise in the surcharge is likely to trigger debate about the cost to consumers of [Berlin]**Berlin**'s energy revolution, a drastic energy policy reversal triggered by the 2011 [Fukushima, Fukushima]**Fukushima** nuclear plant disaster in [Japan]**Japan**. Known as the [--NME--]**Energiewende**, the shift to a sustainable energy supply based on renewable energies and the phasing out of nuclear energy by 2022 has evolved into one of th ...

- Renewable energy statistics [0.243245864581918]
- The EU in the world - energy [0.24015135647884356]
- Sustainable development - Climate change and energy [0.23759774864890335]
- Energy introduced [0.2198400006503435]
- Energy production and imports [0.21531765839384287]
- Electricity and natural gas price statistics [0.19249366143808658]
- Electricity production, consumption and market overview [0.19075948414655808]
- European energy statistics system [0.18545355009532918]
- Consumption of energy [0.18400915935299647]
- Europe 2020 headline indicators [0.174514744551757]

Figure 4.5: **Overview of the LILIANA exploratory interface.**

3. **Disambiguation result box.** The output in the upper right pane shows for each mention (in blue), the assigned entity as a link. The links point to the corresponding Wikipedia articles. Alternatively, they could point to the YAGO ontology entries, or any comparable knowledge source.

4. **Search result box.** Finally, in the lower right pane links to the top ranked statistical reports are shown. In order to help the user finding the most appropriate article, the title of the statistics article and the computed confidence score based on the selected linking method are shown.

4.3.7 Demonstration Scenario

Assume, a user reads a news article on *Germany's switch to renewable energy* and is interested in finding supporting information on Eurostat. The user selects the text and copies it into the input box in LILIANA interface. By default, LILIANA runs with Voting+TFIDF scoring. This can be changed by clicking on one of the buttons with the desirable scoring method.

After submitting the query, the disambiguation result panel holds the input text with highlighted YAGO entities. *RWE* and *Vattenfall* are the detected mentions of the German energy companies.

The result of the live linking is shown in the search result panel. In this case, LILIANA points the user to the article on *Renewable energy statistics* at the top rank.

Another way of interacting with LILIANA is to activate it through the browser plug-in. In the example shown in the screenshot, the user has selected a text fragment of a news article dealing with *Rising Energy Prices – Germans Grow Wary of Switch to Renewables.* Upon clicking on the right mouse button the user is able to select the option *Link to Online Statistics.* This option directs him to our disambiguation and link recommendation server.

The LILIANA browser plug-in is available for download at `https://addons.mozilla.org/de/firefox/addon/liliana-linking/`. LILIANA recommendation server is available at `https://d5gate.ag5.mpi-sb.mpg.de/webliliana/`.

4.4 Application: SESAME

SESAME addresses the use case of finding background information in Wikipedia for numerical data such as observations from statistical tables published by Eurostat. The Eurostat tables are two-dimensional matrices with rows being indexed by countries and columns by years.

Definition 4.2. A *statistical observation O* is a cell in this matrix. In the Figure 4.6, the highlighted cell is an observation. Together with the table title, this observation forms the triple:

$$O = \langle 2015, Germany, Asylum\, and\, first\, time\, asylum\, applicants \rangle.$$

This triple is to be understood as a query to Wikipedia for retrieving related pages. However, the captions of statistical tables are limited in the vocabulary and use very specific terms. For example, *First time asylum applicants* or *Primary production of renewable energy.* This limits the effectiveness of plain keyword search. Instead, within the SESAME system we propose a method for semantic interlinking.

Given an observation from a statistical table and a collection of Wikipedia articles organized in a link graph, we retrieve a ranked list of related Wikipedia articles in the following steps:

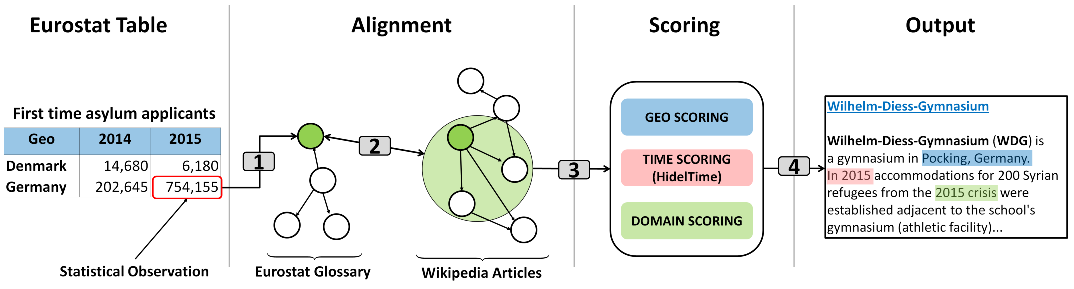
1. we map each statistical observation O onto a set of Eurostat glossary terms;

2. we compute an alignment between the link graph of the Eurostat glossary terms and the link graph of Wikipedia;

3. additionally, we contextualize each Wikipedia article in the collection with the mentions of location and time;

4. using the alignment constructed in the Step 2, we are able to interpret the Eurostat glossary pages as Wikipedia articles. We call these articles *domain pages*. Each article in the Wikipedia collection is scored according to its proximity to the domain pages, location and time of the input observation O.

Step 2 is our main contribution within the SESAME system. We first give a formal definition of the computational model and subsequently describe the algorithmic steps in detail.



Geo	2014	2015
Denmark	14,680	6,180
Germany	202,645	754,155

Wilhelm-Diess-Gymnasium

Wilhelm-Diess-Gymnasium (**WDG**) is a gymnasium in Pocking, Germany. In 2015 accommodations for 200 Syrian refugees from the 2015 crisis were established adjacent to the school's gymnasium (athletic facility)...

Figure 4.6: **Overview of the SESAME pipeline.**

4.4.1 Computational Model

Definition 2.3 serves as a basis for SESAME's computational model. Consider two link graphs G_1 and G_2 and the sets of their instances $I(G_1)$ and $I(G_2)$, respectively. There are directed edges (i-i type) for connecting instances (nodes) of both link graphs.

Analogous to the LILIANA setting, the overlap of instances between the two input graphs is very low. Formally, the task outlined in Step 2 of the SESAME pipeline is defined as follows.

Problem Definition 4.2. Let G_1 and G_2 be two link graphs which have a very small number of overlapping instances. Assume, $\mathcal{A}'$ is an alignment between them: $\mathcal{A}' = \{\langle i_1, i_2, conf \rangle : i_1 \in I(G_1), i_2 \in I(G_2), conf \in \mathbb{R}\}$. Our goal is to find an alignment $\mathcal{A} \supseteq \mathcal{A}'$, which extends the "trivial" alignment $\mathcal{A}'$. That is, for each instance $i_1 \in I(G_1)$ which has a trivial match $i_2 \in I(G_2)$, find a list of most relevant counterparts $i_3 \dots i_n \in I(G_2)$ using the link structure of G_2.

As G_1 SESAME gets supplied the link graph of the Eurostat glossary terms and G_2 is the link graph of Wikipedia articles.

4.4.2 Mapping Tables to Glossary Terms (Step 1)

The goal of this step is to annotate each table with a set of simpler terms which describe the topic of the table. For instance, the table *First time asylum applicants* can be well described with the terms *Dublin regulations* or *Right of Asylum.*

Eurostat statistical reports contain links to the dedicated glossary section, which comprises concepts of that kind. For illustration, we give an excerpt of a statistical report:

> **Asylum quarterly report**[5]
>
> "The number of first time asylum applicants in the EU-28 decreased by -15% in the third quarter of 2016 ...(Table 1)."

Here, *first time asylum applicants* points to the glossary page. From the co-occurrences of the tables and the glossary terms in the reports a set of description terms is built for each table. These description terms are propagated to all the observations (cells) in the table.

4.4.3 Alignment Construction (Step 2)

The objective of the second step is twofold:

1. to be able to interpret the glossary annotations as Wikipedia pages, and

[5] http://ec.europa.eu/eurostat/statistics-explained/index.php/Asylum_quarterly_report accessed on 27.06.2017

Eurostat Glossary Term	**Corresponding Wikipedia Page**
Refugee	Refugee Refugees of the Syrian Civil War Refugee Camp Afgan Refugees
Young people neither in employment nor in education and training (NEET)	NEET Youth unemployment Youth unemployment in Italy Unemployment Jobseeker's Allowance Refusal of Work
Milk production	Dairy Farming Milk Dairy Dairy Cattle Milk Quota

Table 4.1: **Some examples of anchor alignments discovered by SESAME.**

2. to discover a comprehensive set of relevant Wikipedia pages for the input observation based on the glossary annotations.

For this, we devise a two-step procedure.

Anchor Alignment

To detect "trivial" matches between the glossary entries and Wikipedia pages, we follow the search engine-based method and use each glossary term g to query Wikipedia.

Definition 4.3. Formally, we define the *search engine-based similarity* between two instances i_1 and i_2 if i_2 appears within the top-k search results for i_1.

That is, the initial mapping $\mathcal{A}'$ contains triples $\langle g, w, 1 \rangle$, where g is a glossary term and w is a Wikipedia article returned within the top-10 search results. Table 4.1 illustrates some of the anchor alignments.

In contrast to the LILIANA matching strategy, employing search engine allows for detecting complex matches such as *Refugee* and *Dublin Regulations.*

Given $\mathcal{A}'$, we are now able to interpret the glossary annotations of a table as Wikipedia pages. We call them *domain pages* as they describe the topic of the table.

To focus onto semantically coherent table-domain page assignments, we run the ACROSS reasoner prohibiting the following types of alignment:

1. a table is associated with a pair of non-correlating Wikipedia articles. A mapping of a table to both domain pages - *Renewable Energy* and *Population* - is penalized. Two Wikipedia articles are said to be non-correlating, if the Jaccard similarity coefficient over their outgoing links is below a predefined threshold τ.

2. a Wikipedia domain page annotates a pair of non-correlating tables. Two tables correlate positively, if they are mentioned in the reports of the same category. Both tables, *Asylum applicants* and *First residents permits*, belong to the category *Asylum and migration* and, thus, correlate.

Alignment Extension

The domain pages are usually abstract. A large number of pages about events, people or organizations relevant for the input table remain undiscovered. To extend $\mathcal{A}'$, we include all Wikipedia pages semantically related to the domain pages.

Definition 4.4. Let $\langle i_1, i_2, conf\rangle \in \mathcal{A}'$ be a pair of aligned instances, where i_1 is a glossary term and i_2 is a Wikipedia page. We expand this alignment by mapping i_1 onto all articles i'_2 of the Wikipedia graph G_2 strongly related to the article i_2. The *relatedness of two Wikipedia articles* is defined as the Jaccard coefficient of their outgoing links:

$$sim(i_1, i'_2) = \frac{|out(i_2) \cap out(i'_2)|}{|out(i_2) \cup out(i'_2)|} \tag{4.5}$$

Figure 4.2 illustrates this type of extension.

Based on this definition and the set of Wikipedia domain pages $D = \{d_1 \ldots d_n\}$ for an input table t, all relevant Wikipedia pages w for t satisfy the property:

$$rel_d(t, w) = \max_{d \in D} Jaccard(w, d) \geq \theta \tag{4.6}$$

where θ is a predefined threshold. Collectively, we refer to the Wikipedia pages relevant for t as W_t.

4.4.4 Contextualization of Wikipedia Articles (Step 3)

As defined above, a statistical observation O is a cell in a table t indexed by location and time. Thus, we contextualize each related Wikipedia page $w \in W_t$ with these two coordinates:

1. **Temporal mentions.** Using a temporal tagger, we determine all temporal expressions mentioned in w. Due to the largely narrative structure of Wikipedia, we use the domain-sensitive temporal tagger HeidelTime [94] with its narrative normalization strategy to correctly normalize not only explicit dates (e.g., *April 2002*), but also relative and underspecified expressions (e.g., *one month later* and *April*, respectively). As the temporal tag of an observation is always at year granularity, all extracted date expressions of finer granularities (e.g., *April 2002*) are mapped to the respective year (e.g., *2002*) and coarser expressions are ignored (e.g., *20th century*). Thus, each article w is associated with a multiset of year references.

2. **Location mentions.** In order to derive the set of location mentions of w, we consider all the outgoing links and treat them as entities. Using the YAGO ontology, each entity is resolved to a semantic type and only those mapped to *yagoGeoEntity* are selected. Since the geo tag of the table observation is coarse-grained and is always a country, all location mentions in w are mapped to countries via the *locatedIn* relation. Thus, both locations *Berlin* and *Black Forest* are converted to *Germany.*

4.4.5 Ranking Wikipedia Articles (Step 4)

To rank related pages W_t, we introduce a scoring scheme. It computes two proximity measures for each Wikipedia article $w \in W_t$ with respect to the input observation O. The article w is said to be relevant for O if:

1. w is semantically related to the location of O. Let G be the set of links pointing from w to any geo entity, and G' be the set of outgoing links to the entities associated with O's location. The *location relevance* of page w is

$$rel_g(w) = \frac{|G'|}{|G|}$$

2. w is relevant to the time. Let Y' be the number of year mentions related to the year of the observation and Y be the total number of temporal expressions. The *time relevance* for page w is

$$rel_t(w) = \frac{Y'}{Y}$$

The final relevance score of w considers the two previously introduced proximity measures together with the domain relevance $rel_d(t, w)$ by a linear combination of their weights as follows:

$$rel(w, O) = \alpha \cdot rel_d(t, w) + \beta \cdot rel_g(w) + \gamma \cdot rel_t(w) \qquad (4.7)$$

The time relevance can be further adjusted by considering the creation date of w. When looking for the recently emerged entities and events, the set of relevant articles can be focused onto those, which were created in the year of the observation. However, since the Wikipedia history begins in January 2001 constraining to earlier years is not possible. Thus, using the creation time stamp of w is left to the user as an option, rather than a part of the scoring scheme.

4.4.6 Data

We have crawled the statistical reports from the Statistics Explained portal of Eurostat[6] in March 2016. The data contains 2,472 reports, 1,990 glossary terms and 557 categories. The tables with numerical statistics are taken from the Eurostat Database[7]. In total, there are 2,398 tables which are mentioned in statistical reports. A minor fraction of the tables have time series of monthly or quarterly granularity. For demonstration purposes and the sake of comparability, only tables with yearly statistics are considered.

To compute the domain similarity of the pages, we use the static link graph derived from the English Wikipedia dump as of June 1, 2016. The revision history is parsed from the meta-history dumps of the same date and captures user activities starting from January 16, 2001. The page view data is retrieved from `http://stats.grok.se`.

[6] `http://ec.europa.eu/eurostat/statistics-explained/index.php/Main_Page` accessed on 27.06.2017

[7] `http://ec.europa.eu/eurostat/data/database` accessed on 27.06.2017

4.4.7 Implementation

The SESAME interface is built using the Ace admin template[8], which is based on the Bootstrap framework and JQuery. On the server side, SESAME is written in Java and runs on a Tomcat server.

SESAME precomputes and stores the following data in a PostgreSQL database:

- Wikipedia anchors for statistical tables;
- the Wikipedia link graph;
- annotation of Wikipedia articles with YAGO geo entities; and
- temporal mentions found in Wikipedia articles by HeidelTime.

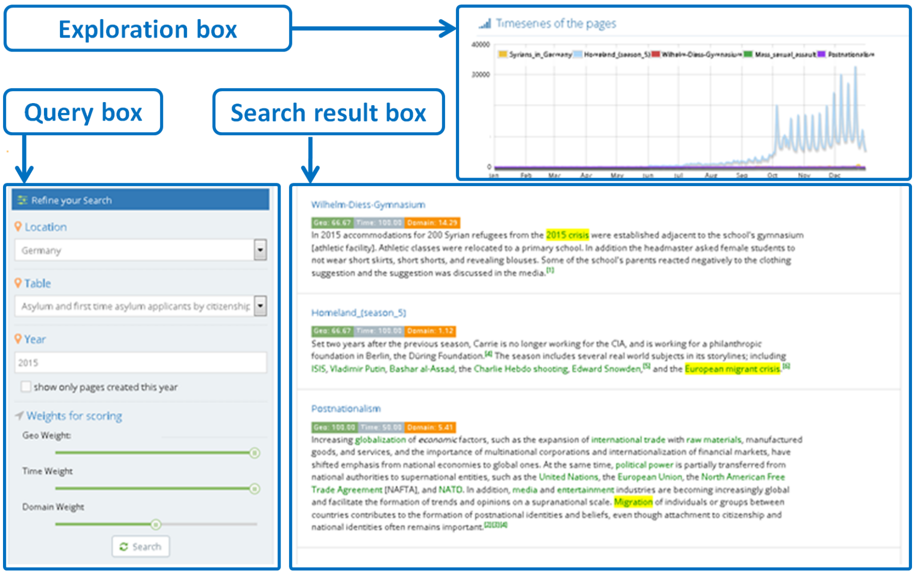
Through the Web interface, a user submits a query (table, location, year, weights for the three similarity measures) to the back-end engine. Each of the three scorers (see Fig. 4.6) retrieves relevant documents. To this end, the scores are aggregated using according to the user preferences. The scorers also return the "provenance" - the links through which the articles were considered to be relevant. These include location mentions and links to Wikipedia anchor pages. The server returns the ranked list of Wikipedia articles and the visualization interface renders the search results with provenance highlighting.

The SESAME user interface is divided into three parts:

1. **Query box.** The form on the left side serves for issuing the query. The user specifies the scope of numerical statistics to be contextualized by selecting a table, location and the year. Three types of weights - location, time and domain - may be adjusted by the user and are used for producing the aggregated score for relevant Wikipedia articles.
2. **Search result box.** After submitting the query, ranked search results are displayed on the right panel. Each article is represented in an expandable box. The box contains text snippets, where the anchors and relevant location mentions are highlighted.
3. **Exploration box.** The top panel provides two containers for further exploration. The left top box displays the selected Eurostat table and allows users to inspect the numerical statistics itself. The right box serves for highlighting the dynamics of the corresponding Wikipedia pages. To

[8] `http://ace.jeka.by/` accessed on 27.06.2017

this end, it allows three types of user activities to be visualized: number of page views, number of editors, and number of page revisions aggregated per day for the query year.



Refine your Search

Location: Germany

Table: Asylum and first time asylum applicants by citizenship

Year: 2015

□ show only pages created this year

Weights for scoring

Geo Weight:

Time Weight

Domain Weight

Search

Wilhelm-Diess-Gymnasium

Geo: 66.67 Time: 100.00 Domain: 14.29

In 2015 accommodations for 200 Syrian refugees from the 2015 crisis were established adjacent to the school's gymnasium (athletic facility). Athletic classes were relocated to a primary school. In addition the headmaster asked female students to not wear short skirts, short shorts, and revealing blouses. Some of the school's parents reacted negatively to the clothing suggestion and the suggestion was discussed in the media.[1]

Homeland_(season_5)

Geo: 66.67 Time: 100.00 Domain: 1.12

Set two years after the previous season, Carrie is no longer working for the CIA, and is working for a philanthropic foundation in Berlin, the Düring Foundation.[4] The season includes several real world subjects in its storylines; including ISIS, Vladimir Putin, Bashar al-Assad, the Charlie Hebdo shooting, Edward Snowden,[5] and the European migrant crisis.[6]

Postnationalism

Geo: 100.00 Time: 50.00 Domain: 5.41

Increasing globalization of *economic* factors, such as the expansion of international trade with raw materials, manufactured goods, and services, and the importance of multinational corporations and internationalization of financial markets, have shifted emphasis from national economies to global ones. At the same time, political power is partially transferred from national authorities to supernational entities, such as the United Nations, the European Union, the North American Free Trade Agreement (NAFTA), and NATO. In addition, media and entertainment industries are becoming increasingly global and facilitate the formation of trends and opinions on a supranational scale. Migration of individuals or groups between countries contributes to the formation of postnational identities and beliefs, even though attachment to citizenship and national identities often remains important.[2][3][4]

Figure 4.7: **Overview of the SESAME exploratory interface.**

4.4.8 Demonstration Scenario

Our demo covers an entire SESAME walk-through. Suppose, a user is interested in finding background information for the table *Asylum and first time asylum applicants*, for Germany in the year 2015. To build the query, the user navigates to the query box. The table and country are selected from the drop-down list and the year is typed into the corresponding field. Further, the feature weights can be adjusted with sliders according to the user's preferences.

Viewing the table data in the right exploration box reveals, that the number of asylum applications has risen from 202,645 in 2014 to 476,510 in the year 2015.

After submitting the query, the ranked Wikipedia articles are displayed in the search result box. Articles *Wilhelm-Diess-Gymnasium* and *Homeland (season 5)* are the top-ranked related pages, if location and time weight are set to 1.0 and the domain weight is 0.05:

Wilhelm-Diess-Gymnasium

"In 2015 accommodations for 200 Syrian refugees from the 2015 [migrant] crisis were established adjacent to the school's gymnasium (athletic facility)."

> **Homeland (season 5)**
>
> "The season includes several real world subjects in its storylines; including ISIS, Vladimir Putin, ... and the European migrant crisis."

The top right plot visualizes the page views and edit dynamics of the detected relevant articles.

Prominent articles such as *List of migrant vessel incidents in the Mediterranean Sea* attract considerable amount of community attention. On September 5th, 2015 this page had 1708 views. However, the long-tail article about Wilhelm-Diess-Gymnasium has a small number of editors and is also underexplored by Wikipedia users. It had moderate number of page views in 2015, with its maximum of 31 on September 9th and with at most 14 views per day until the end of 2015. Sesame is able to detect related pages independent of their prominence and unlock the hidden information from both, popular and long-tail articles.

The screencast of SESAME is available at `https://youtu.be/H2TiSTwqUhU`.

The live demo of SESAME is available at `https://gate.d5.mpi-inf.mpg.de/sesame/`.

4.4.9 Experimental Evaluation

Ground Truth Construction

To evaluate the quality of SESAME alignments, we have randomly selected 20 observations. We run all methods under observation and produce a pool of alignments per observation. Recommended Wikipedia pages are annotated by human judges either as *relevant* or *not relevant*. Our evaluation instructions stated that a page is considered to be relevant, if it is topically related to the table, as well as to the location and the time. In total, 510 alignments were evaluated.

Methods

We have the following methods under comparison:

1. **SESAME** with α, β, γ manually set to 0.5 (Formula 4.7). In these runs, the time relevance is controlled solely by parameter γ. It means, that temporal relevance is determined solely from temporal expressions contained on a Wikipedia page.

2. **SESAME+time** additionally considers the creation date of Wikipedia articles (equivalent to checking *show only pages created this year* in the GUI).

3. **Search engine** is our baseline approach. We have chosen two popular search engines - Google and Bing. To find relevant pages, we explicitly set the desired domain as *site:en.wikipedia.org* and formulate the queries as table name, location and year separated by a white space. A sample query might look as:

   > *site:en.wikipedia.org Asylum and first time asylum applicants Germany 2015.*

4. Further, in the **Search engine+time** mode, we exploit the time settings utility provided by the search engines by additionally specifying the year of document creation (e.g., in Bing by setting *Date* → *Custom range* to 01.01.2015 to 31.12.2015 for the year 2015).

Measures

We report on four quality measures. Formulas for computing some of them were given in Section 3.9.3. Let O be the input statistical observation, $rel(O, i)$ be the relevance judgement for the recommended Wikipedia article at rank i and $Cand(O)$ be the ranked list of Wikipedia articles returned for the statistical observation. The reported measures are:

1. **Mean Reciprocal Rank**;

2. **Mean Average Precision** with cut-off at rank 10;

3. **Precision@10:** the precision for the top-10 ranks only:

$$prec@10(Cand(O)) = \frac{|\{i : rel(O, i) = relevant \ \& \ i <= 10\}|}{\min(|Cand(O)|, 10)} \tag{4.8}$$

4. **Success@1** refers to the portion of tables for which a relevant Wikipedia article was found at rank 1.

Method	**MRR**	**MAP@10**	**Prec@10**	**Succ@1**
SESAME	0.69	0.44	0.48	0.55
SESAME + time	0.54	0.17	0.27	0.41
Google	0.24	0.18	0.23	0.17
Google + time	0.20	0.03	0.13	0.0
Bing	0.41	0.11	0.23	0.22
Bing + time	0.14	0.02	0.05	0.0

Table 4.2: **Experimental results of SESAME.**

Results

The experimental results are given in Table 4.2. Table 4.3 provides anecdotal search results returned by SESAME, Google and Bing.

SESAME finds related Wikipedia pages with fairly high MRR and MAP@10 values. In contrast to SESAME, which treats locations as entities and also utilizes an underlying knowledge base to capture all locations belonging to a country by considering *locatedIn* relations, the search engines were not able to properly find pages related to the locations and time, treating these terms rather as keywords. This resulted in finding many general pages, which are location- and time-neutral (e.g., *List of countries by public debt* or *Economy of the Czech Republic*).

Moreover, the table names are lengthy and are rather hard to deal with for a search engine. SESAME resolves this shortcoming by "reformulating" table names as a set of domain pages. All the systems under consideration performed with lower MAP values when the creation date of Wikipedia pages was constrained (+ time option). This can be explained by the following observations: once an event has happened, there are many already existing Wikipedia pages which get updated (such as *Wilhelm-Diess-Gymnasium* in conjunction with the migrant crisis); pages for re-occurring events are created prior to their planned dates (already existing page *Olympics 2020* is a good illustration). By limiting the page creation date, all the systems loose a large portion of relevant results.

Query	SESAME	Search Engine
Marriage Indicators, UK, 2010	**Islamic Sharia Council** The Islamic Sharia Council (ISC) is a British organisation that provides legal rulings and advice to Muslims in accordance with its interpretation of Islamic Sharia based on the four Sunni schools of thought. It primarily handles cases of marriage and divorce and, to a lesser extent business and finance. The council has no legal authority in the United Kingdom, and cannot enforce any penalties; many Muslims would appear voluntarily to accept the rulings made by the ISC. A rival service, the Muslim Arbitration Tribunal, was founded in 2007 by followers of the Barelvi school of South Asian Islam, is reportedly "less strict than the Deobandis" and as of 2010 offered dispute resolution in half a dozen British cities.	**Google:** **List of countries by age at first marriage** **Bing:** **List of countries by age at first marriage**
Gross Nutrient Balance, Greece, 2009	**Tephritid Workers Database** A group of scientists involved in tephritid fruit fly research and management launched the Tephritid Workers Database on May 2004, with the support of the Insect Pest Control Section of the Joint Food and Agriculture Organization /International Atomic Energy Agency programme. Initiated in 1982 at the First International Symposium held in Athens, the quadrennial fruit fly symposium for the international fruit fly workers is being well established now with a large number of scientists from all over the world attending the symposium. This newsletter publication was interrupted in 1992 and then resumed in an electronic format since 2009.	**Google:** **Sustainable agriculture** **Bing:** **List of countries by public debt**
At-risk-of-poverty Rate, Czech Republic, 2005	**Decade of Roma Inclusion** The initiative was launched in 2005, with the Decade of Roma Inclusion running from 2005 to 2015, and was the first multinational project in Europe to actively enhance the lives of Roma. The 12 countries taking part in the Decade of Roma Inclusion were: Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Hungary, Macedonia, Montenegro, Romania, Serbia, Slovakia, and Spain. The governments of the above countries have committed to closing the gap in welfare and living conditions between the Roma and non-Roma populations, as well as putting an end to the cycle of poverty and exclusion that many Roma find themselves in.	**Google:** **Poverty in Germany** **Bing:** **Economy of the Czech Republic**

Table 4.3: **Anecdotal top-1 search results produced by SESAME, Google and Bing. SESAME outputs short annotations explaining the alignment decision. It highlights the relevant parts of a Wikipedia page as follows: yellow – relevant domain links, green – relevant location links, and gray – relevant temporal mentions. For Google and Bing only document titles are listed.**

4.5 Related Work

We now discuss how LILIANA and SESAME are related to several areas of previous research.

The most relevant related work for both, the LILIANA and the SESAME systems, embraces the field of **linking Wikipedia to external sources**. Mishra et al. [69] study the problem of retrieving relevant news for Wikipedia excerpts based on the contained entity, location and time mentions, as well as textual overlap. Nanni et al. [71] build a system for discovering entities associated with events based on their co-occurrence in Wikipedia articles and the proximity within the Wikipedia link graph. Additionally, they retrieve passages that explain the relationships between the entities and the events. Spitz et al. [91] introduce a model for event retrieval in a large-scale collection of documents. The entire collection is converted to a co-occurrence graph of entities. This graph is subsequently used in the retrieval step for obtaining most related entities for the input query. The reviewed works are concerning about perfectly matching entities mentioned in the query to the entities detected in the document collection. Analogously to the system developed by Spitz et al., SESAME targets extending the equivalence matches with closely related ones. SESAME, in accordance to the previous systems, formulates a query as a $\langle concepts, time, location \rangle$ triple. However, SESAME constructs the query given a numerical table and a collection of documents referring to it.

Semantic search and entity retrieval have been studied by Balog et al. [5], Nie et al. [75] and Elbassuoni et al. [27]. They require as input either a keyword or a structured query. Balog et al. devise a probabilistic framework for entity search based on the terms, categories or related entities. The entity ranking scheme includes co-occurrences of terms and entities in a background corpus of documents and Wikipedia entity categorization. STICS implemented by Hoffart et al. [43] operates with semantic categories for retrieving a set of relevant news articles. LILIANA's retrieval approach is similar in spirit to these works. However, the novelty is in constructing an alignment between the topic system of Eurostat and the YAGO type system.

Semantic linking of tables. Bhagavatula et al.[8] and Limaye et al. [58] aim at identifying entities mentioned in Web tables and the relations between them. The lack of a common scheme and high ambiguity of entity mentions are primarily addressed problems in these works. SESAME, in contrast, deals with tables having a unified schema. The focus of SESAME is on identifying related documents from Wikipedia using time and location information as a part of ranking procedure.

Research in **credibility assessment** has recently seen increasing interest as the vast volume of claims are made available on social media. T-Verifier developed by Li et al. [57] is a system for automated credibility assessment for doubtful statements. The core idea of the system is in computing the ratio between the portion of Web documents containing a doubtful claim compared to those containing the alternative claims. Castillo et al. [16] target assessing credibility of content published on Twitter. They employ an extensive list of features and train a supervised classifier to assign a trustfulness score to each Twitter message. Popat et al. [82] developed a framework for verifying long-tail facts and providing explanations for the credibility decision. The credibility decision is inferred based on the language used for reporting the claim, the reliability of the source and its stance towards the claim. Linking to or from a high authoritative source is the main concern of LILIANA and SESAME. Automated separation of true statements from rumors is, however, beyond the scope of our work. The techniques proposed in this chapter can be used for focused information extraction for credibility assessment, though.

4.6 Summary

This chapter presented two systems – LILIANA and SESAME – for aligning taxonomies and link graphs given their high structural heterogeneities. The core idea of both systems is to find the small set of categories or instances common to both taxonomies or link graphs and to expand this alignment utilizing structure-level features. We demonstrated two applications which vitally depend on the ability to cross the borders between structurally different knowledge repositories. Both, LILIANA and SESAME, raise data analytics to the entity-level and help discovering mutual dependencies between Web content and statistics published by government. This is a potential gold mine for researchers like sociologists, politologists, media and market analysts.

Chapter 5

LAIKA: Alignment of Dynamic Knowledge Repositories

Web knowledge repositories are inherently dynamic and large. Once an alignment is computed, it has to be monitored to reflect changes in the interlinked repositories. One approach is to regularly rerun the entire alignment pipeline on fresh snapshots of knowledge repositories. This solution comes with a considerable overhead when knowledge repositories are large (such as Wikipedia totalling more than 41 million pages for the English edition as of 22.03.2017) and with frequent updates. Another drawback is re-discovering already existing alignments. We propose a method that addresses detection of missing alignments relying on the existing correspondences between two link graphs. This method is the core of the LAIKA system, our contribution to supporting alignment curation for dynamic knowledge repositories.

When the alignment is constructed, it can further be used to help one knowledge repository to learn from the aligned (parallel) repository. It means, to discover more relations between instances and categories relying on the external link structure. This task is referred to as knowledge repository curation and is addressed in LAIKA as well. More specifically, we target finding missing article-article and article-category links within one Wikipedia edition using a Wikipedia of a different language as evidence.

Methods presented in this chapter are generalizable to any graph-structured knowledge repositories, for which a partial alignment is given. For illustrative purposes, we consider LAIKA's application to the problem of curating inter- and intra-language links for two Wikipedia editions.

5.1 Motivation

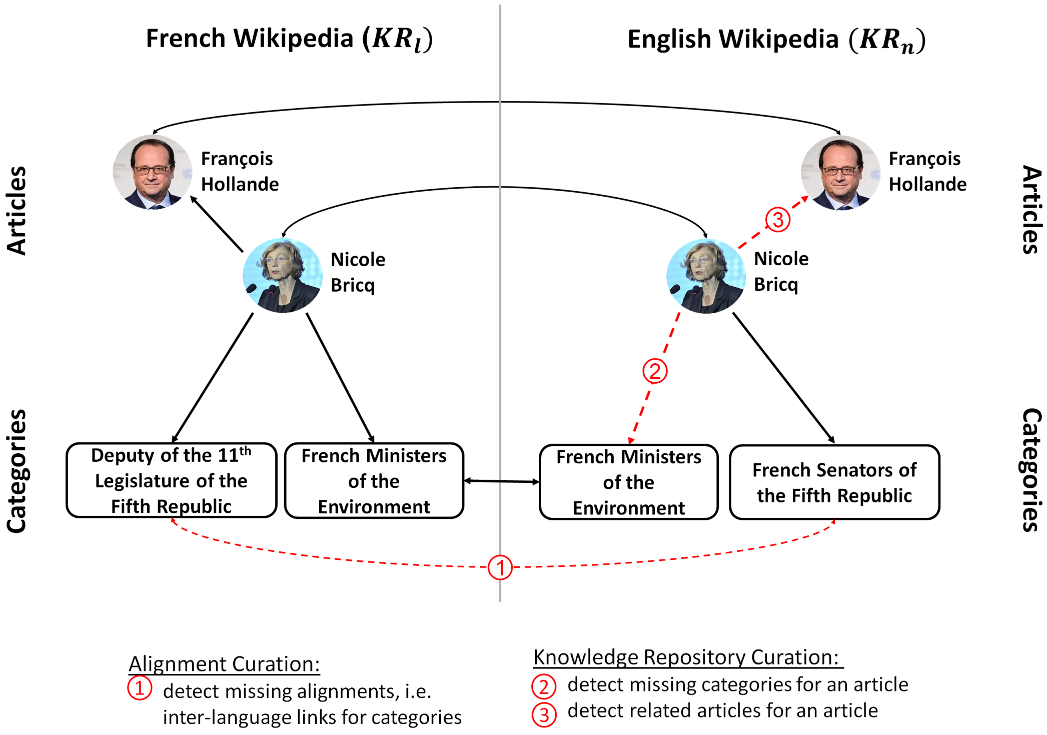

French Wikipedia (KR_l)
English Wikipedia (KR_n)
Articles
Categories
François Hollande
Nicole Bricq
François Hollande
Nicole Bricq
Deputy of the 11th Legislature of the Fifth Republic
French Ministers of the Environment
French Ministers of the Environment
French Senators of the Fifth Republic
Articles
Categories
1
2
3
Alignment Curation:
1 detect missing alignments, i.e. inter-language links for categories
Knowledge Repository Curation:
2 detect missing categories for an article
3 detect related articles for an article

Figure 5.1: **An example of (a) an incomplete alignment between the French and the English Wikipedia editions and (b) of the missing links within the English Wikipedia. For clarity, the titles of all articles and categories are presented in English. Black links within and across the two Wikipedias are existing. The red links are missing. The cross-language links between categories are understood as an existing albeit incomplete alignment.**

Knowledge-sharing communities like Wikipedia keep growing at impressive rates[1]. Ideally, whenever an article or a category is added into a knowledge repository, it immediately should be linked to the equivalent pages or categories in a counterpart KR. Consider the example presented in Figure 5.1. The French category *Deputy of the 11th Legislature of the Fifth Republic* is equivalent to the English category *French Senators of the Fifth Republic*. However, the inter-language link between these categories is missing. Editors of large communities such as Wikipedia are overwhelmed with the amount of information which is

[1] `http://stats.wikimedia.org` accessed on 27.06.2017

added to different language editions of Wikipedia every day. Manually tracing missing inter-language links between several Wikipedia editions is a very laborious task. This motivates the need in a machinery for recommending most related translations for a Wikipedia category in a target language. Approaches for missing link detection address this kind of problems. However, they either use an expensive optimization technique [38, 98] which make them inapplicable for large-scale graphs or discriminate low-connectivity nodes [51]. These considerations shape the requirements to the recommendation procedure:

- to perform incremental alignment updates;
- to be able to detect and score candidate counterparts in on-line manner and with high accuracy.

We refer to this task as *alignment curation*.

Another aspect of knowledge repositories is their difference in dynamics. As a concrete example, consider announcing the ministers of the French government on the 16th of May 2012. Nicole Bricq was appointed Minister of Ecology and Sustainable Development. Consider now the Wikipedia pages about Nicole Bricq in French and English Wikipedia a full week after the original event. The corresponding link graphs are illustrated in Figure 5.1. The French version of the Nicole Bricq's page is an extensive biography. In May 2012 it was linking to many relevant concepts and entities, such as the French president François Hollande. Her English Wikipedia page merely consisted of a single short paragraph with the comment that it needs to be expanded. Despite the detailed article in the French Wikipedia and the media reports about the new government, the English-speaking Wikipedia community was not able to detect missing facts for the Nicole Bricq's page.

In this scenario, the idea of knowledge repository curation would be to give recommendations to the non-French communities about pages that should be expanded or categories into which new or expanded articles should be placed. For the illustration, refer to the missing links denoted as 2 and 3 in Figure 5.1. Such recommendations for additional related articles and categories should be generated in an automated manner by analyzing several Wikipedia editions across languages and ideally considering online news that mention Wikipedia entities in at least one language (in the form of hyperlinks or linked-data formats like RDFa[2] statements).

We call this type of maintenance *knowledge repository curation*.

[2] `http://www.w3c.org/TR/xhtml-rdfa-primer/` accessed on 27.06.2017

5.2 Contribution

The LAIKA system addresses the outlined problems of alignment and knowledge repository curation and casts them into link recommendation tasks. For recommending links within and across Wikipedia editions, we use link-overlap measures such as weighted Jaccard and random walk techniques such as SimRank [48]. We generalize the notion of SimRank to work with our model of Wiki link graph, introducing a weighted extension of SimRank.

In summary, the novel contributions presented in this chapter are:

- *alignment curation:* we define the type of correspondences for which we construct the alignment and perform incremental updates;
- *knowledge repository curation:* we introduce two types of link recommendation problem based on an aligned counterpart KR;
- *algorithms for detecting missing links:* we develop a suite of efficient algorithms for predicting and ranking missing links within alignment curation and knowledge repository curation problems;
- *experiments with large Wikipedia graphs:* we report on experimental studies with Wikipedia editions in three different languages.

5.3 Computational Model

Consider link graphs of Wikipedia pages (articles and categories) from two language editions as illustrated on Figure 5.1. We refer to these link graphs as KR_l and KR_n for languages l and n respectively.

As described in Section 2.1, $I(KR_l)$ denotes the set of articles of Wikipedia edition KR_l, and $C(KR_l)$ are categories of KR_l. Within one language edition, LAIKA regards two types of edges:

- *i-c* edges for connecting articles to their categories;
- *i-i* edges for interlinking articles.

For KR_n, the corresponding graph is constructed in similar way.

Wikipedia editions deliver an extensive inter-linkage between equivalent articles or categories across languages. This can be understood as a partial alignment between two Wikipedia editions KR_l and KR_n.

Analogous to the notation used in Chapter 3, let $A_{i,j}$ and $A_{c,k}$ refer to the bi-directional inter-language links between articles i and j or categories c and k. We denote the set of all alignments (i.e., all links between KR_l and KR_n) as $\mathcal{A}$. When the type of a node is irrelevant, we simply denote it v.

5.4 Overview of LAIKA

LAIKA addresses two types of curation: (a) ensuring up-to-date alignments between two Wikipedia editions (inter-language links between categories) and (b) capturing missing links from the aligned Wikipedia edition (intra-language links). In this section, we present both curation tasks and outline algorithms for solving them. The following building blocks are common to all algorithms:

1. *Candidate target detection.* For a given start node v we compile a set of potential targets $Cand(v)$ by traversing the graph composed of KR_l, KR_n and $\mathcal{A}$ with a bounded number of hops.

2. *Ranking.* All targets in $Cand(v)$ are ranked according to a chosen similarity measure.

Both stages utilize only the link structure of KR_l and KR_n. Analyzing textual content of articles and categories for aligning and scoring is subject to follow-up research.

5.5 Candidate Target Detection (Step 1)

This section is devoted to the algorithms for finding candidate targets $Cand(v)$ depending on the task (first stage in our algorithmic skeleton). The notation needed for algorithm description is presented in Table 5.1. Ranking functions are considered in the next section.

5.5.1 Alignment Curation

We focus on finding missing inter-language links for categories because we consider them being a bottleneck in cross-lingual data quality. We disregard the simplest and most obvious prediction type: recommending inter-language links between articles, because Wikipedia editions already exhibit very high coverage and accuracy in this regard. For categories, however, the inter-language linkage is much sparser and noisier.

Symbol	**Meaning**
$instances(c)$	Get all instances belonging to category c
$categories(i)$	Get all categories to which an instance i belongs
$related\_articles(i)$	Get all articles connected to i with an outgoing link.
$get\_aligned(v, \mathcal{A})$	Get aligned instances or categories for v in the parallel knowledge repository using $\mathcal{A}$. The function returns $Null$ if there is no alignment for v. Note, that in Wikipedia an article or a category has at most one counterpart.

Table 5.1: **Notation used for describing algorithms for candidate target detection.**

Problem Definition 5.1. Formally, the task of *alignment curation* is: given two knowledge repositories KR_l and KR_n and an existing alignment between them $\mathcal{A}$, find a set of missing alignments $A_{c,k} \in \mathcal{A}'$, where $c \in C(KR_l)$ and $k \in C(KR_n)$ are categories of both knowledge repositories.

When we consider two Wikipedia editions as partially interlinked knowledge repositories, $\mathcal{A}$ is to be understood as the already existing cross-lingual links between these editions.

Algorithm 5.1 describes how a set of potential counterparts $Cand(c) \subseteq C(KR_n)$ is found for a source category $c \in C(KR_l)$. Our intuition behind this algorithm is that for two categories to be equivalent counterparts it is necessary (but not sufficient) to have an overlapping set of articles.

Below we discuss the main steps of Algorithm 5.1. They are additionally illustrated in Figure 5.2:

1. Take all articles of the source category c. These are marked as I_c^l.

2. For all the members of I_c^l, find their equivalences using the existing cross-lingual links $\mathcal{A}$. This set we denote as I_c^n.

3. Take all categories in KB_n that contain any article $i^n \in I_c^n$. These are c_1, c_2 and c_3 in the example.

4. Disregard all categories which are already linked to any category in KB_l. We enforce a 1-to-1 alignment in this case to be consistent with the

Algorithm 5.1: Alignment Curation

```
    Input: category $c \in C(KR_l)$
    Output: a set of potential counterparts $Cand(c) \subseteq C(KR_n)$
1   $I_c^l = instances(c)$ ;
    $I_c^n = \emptyset$ ;
2   foreach $i^l \in I_c^l$ do
        $i^n = get\_aligned(i^l, \mathcal{A})$;
        $I_c^n = I_c^n \cup \{i^n\}$;
    $Cand(c) = \emptyset$ ;
3   foreach $i^n \in I_c^n$ do
        $C^n = categories(i^n)$;
        foreach $c^n \in C^n$ do
4           if $get\_aligned(c^n, \mathcal{A}) = Null$ then
                $Cand(c) = Cand(c) \cup \{c^n\}$;
    return $Cand(c)$;
```

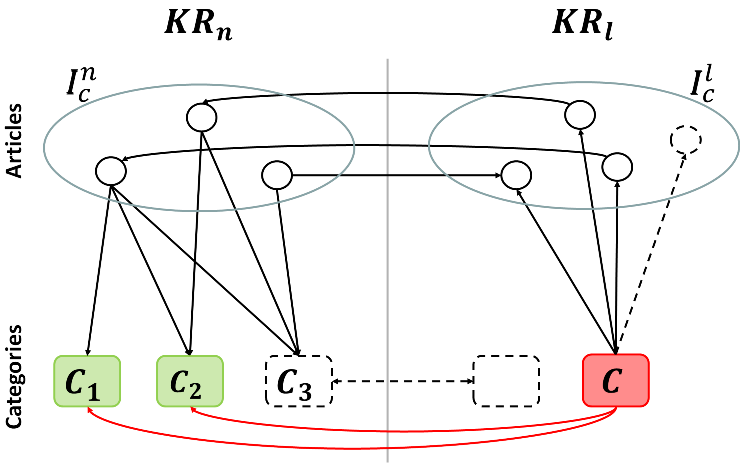


Figure 5.2: **Alignment Curation. The red links denote detected missing alignments. Category C_3 is filtered out since it already has a counterpart in KR_l.**

Wikipedia guidelines. Categories gathered in this way are referred to as candidate targets $Cand(c)$. In the figure, c_3 is already linked to a category in KB_l and, therefore, can not be linked to any other category. Thus the candidate targets for c are $Cand(c) = \{c_1, c_2\}$.

5.5.2 Knowledge Repository Curation

The link recommendation types considered below aim at supporting 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 contents of the new article. Although both recommendation algorithms have input (an article) and output (categories or related articles) in the same language, the point of these recommendations is to utilize the existing alignment $\mathcal{A}$ and contents of a counterpart knowledge repository.

Problem Definition 5.2. Formally, the task of *knowledge repository curation* is: given two knowledge repositories KR_l and KR_n and an existing alignment $\mathcal{A}$ between them, find a set of missing intra-language links of *i-c* and *c-c* type between instances and categories of KR_n based on the link structure of KR_l.

New Categories for an Article

The relevance criteria of category c to article i is that there is a membership relation between the equivalent article and the equivalent category in the aligned knowledge repository. Consider node *Nicole Bricq* in Figure 5.1. It is a member of the category *French Ministers of the Environment* in the French Wikipedia. The goal is to discover all missing categories for the English version of *Nicole Bricq.*

Figure 5.3 and Algorithm 5.2 illustrate how the set of candidate categories $Cand(i^l)$ is found for article $i^l \in I(KR_l)$.

The steps to detect missing article-category links are:

1. Let i^n be the equivalent article for i^l in KR_n.

2. Consider all categories of i^n and

3. map them back to KR_l. Thus, we obtain the set $\{C_1, C_2, C_3\}$.

4. We disregard all the categories which are already assigned to i^l. The resulting set of candidates for the example is $\{C_1, C_3\}$.

Algorithm 5.2: Knowledge repository curation: detecting new categories for an article.

```
Input: article $i^l \in I(KR_l)$
Output: a set of potential related categories $Cand(i^l) \subseteq C(KR_l)$
$i^n = get\_aligned(i^l, \mathcal{A})$ ;
$C^n = categories(i^n)$ ;
$Cand(i^l) = \emptyset$ ;
foreach $c^n \in C^n$ do
    $c^l = get\_aligned(c^n, \mathcal{A})$;
    if $c^l \notin categories(i^n)$ then
        $Cand(i^l) = Cand(i^l) \cup \{c^l\}$;
return $Cand(i^l)$;
```

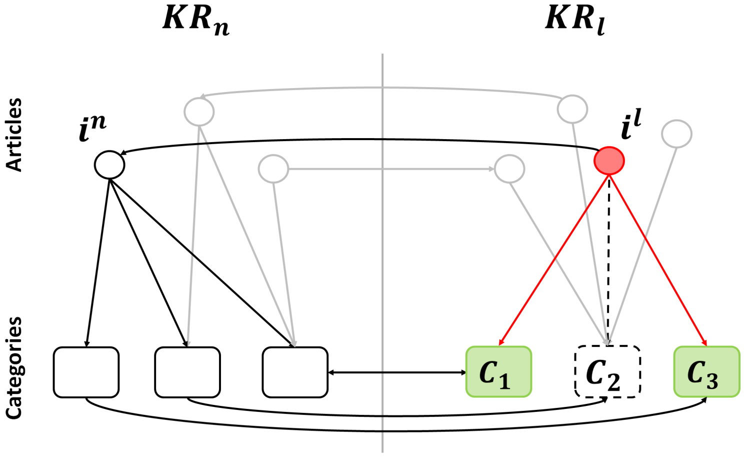


Figure 5.3: **Knowledge repository curation: detecting new categories for an article. Red links connect the source instance to the missing categories. Category C_2 already links to i^l, therefore it is disregarded. The shaded part of the graph is irrelevant for the candidate detection step.**

Related Articles for an Article

It is not rare, that two Wikipedia pages are inter-linked in one language edition and are disconnected in another one. Consider the nodes *Nicole Bricq* and *François Hollande* in Figure 5.1.

To address the problem of that type, we propose an algorithm for transferring the link structure (i-i links) from knowledge repository KR_l onto KR_n for a given source article $i^l \in I(KR_l)$. The procedure is analogous to detection of missing relevant categories described in Algorithm 5.2. The only difference is that instead of considering categories in Step 2, we fetch all articles referenced from i^n. Algorithm 5.3 and Figure 5.4 illustrate the proposed procedure for detecting candidate articles. In this example, j_1 is the only candidate article, since the connection to j_2 already exists in KR_n.

Algorithm 5.3: Knowledge repository curation: detecting related articles for an article.

```
Input: article $i^l \in I(KR_l)$
Output: a set of potential related articles $Cand(i^l) \subseteq I(KR_l)$
$i^n = get\_aligned(i^l, \mathcal{A})$ ;
$I^n = related\_articles(i^n)$ ;
$Cand(i^l) = \emptyset$ ;
foreach $i^n \in I^n$ do
    $j^l = get\_aligned(i^n, \mathcal{A})$;
    if $j^l \notin related\_articles(i^n)$ then
        $Cand(i^l) = Cand(i^l) \cup \{j^l\}$;
return $Cand(i^l)$;
```

5.6 Candidate Scoring Methods (Step 2)

Scoring candidate targets $Cand(v)$ with respect to the source node v is the second building block of LAIKA's algorithmic skeleton. LAIKA's similarity measures are based purely on the graph link structure. This section discusses similarity metrics of two kinds: (a) based on the local neighbourhood of the source and a target node or (b) based on the global link structure of knowledge repository.

5.6.1 Similarity Using Local Structure

Jaccard-based Methods

Use of the local link structure is motivated by its simplicity and suitability for online computations. Let v be a source node and v' a potential target. We denote their direct neighbour sets as V and V' respectively.

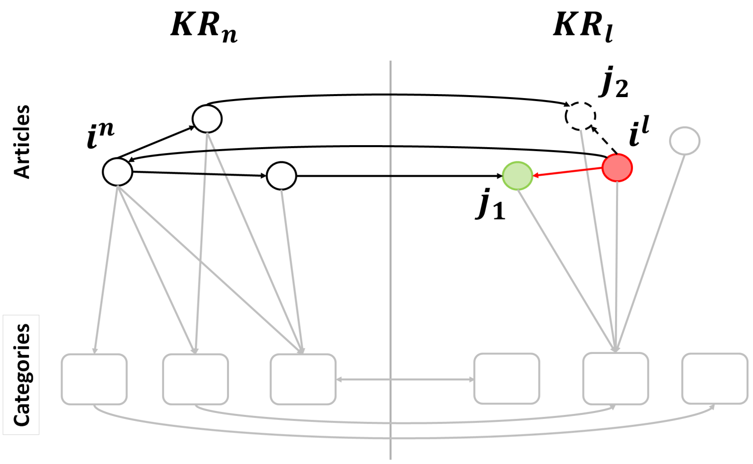


Figure 5.4: **Knowledge repository curation: relevant article detection. The red link denotes the missing connection between i^l and j_1. The shaded part of the graph does not participate in detecting candidate articles.**

We introduce two similarity measures which are implemented by Jaccard coefficients over the neighbour sets. The standard Jaccard coefficient is defined as:

$$j\text{-}sim(v, v') = j\text{-}sim(V, V') = \frac{|V \cap V'|}{|V \cup V'|} \tag{5.1}$$

Definition 5.1. In addition to it, we use *weighted Jaccard* [35]:

$$wj\text{-}sim(V, V') = \frac{\sum_{v \in V \cap V'} min(w(v, V), w(v, V'))}{\sum_{v \in V \cup V'} max(w(v, V), w(v, V'))} \tag{5.2}$$

where $w(v, V)$ is a weight of a neighbour node v in the set V. Weights of article nodes are proportional to the number of categories to which they belong and weights of category nodes are proportional to the number of articles a category holds.

These two measures can be applied to any of the curation tasks we outlined in the previous section.

Voting Method

Definition 5.2. In addition, we devise a simpler method of overlap-based *voting*:

$$voting(V, V') = |V \cap V'| \tag{5.3}$$

This is a simple heuristic, which is tailored to the alignment curation task – finding missing inter-language links between categories of two Wikipedia editions. The heuristic is to directly compare article sets of a source category c_1 and a set of possible counterpart categories $c'_1 \dots c'_n$, where article sets are restricted to those that have bi-directional inter-language links between two respective Wikipedia editions.

The voting method prefers the target with the largest number of articles shared with the source category. While resulting in unnormalized similarity values, this metric is computationally efficient and serves ranking purposes well. This is illustrated by our experimental results in section 5.7.4.

5.6.2 Similarity Using Global Structure

SimRank

Despite their simplicity, the overlap-based measures fail when the source and a target node do not have common neighbours. In this case, we rely on the global structure of multilingual Wikipedia link graphs to compute similarity between a pair of nodes.

Definition 5.3. *Standard SimRank* [48] is a widely used measure of structural context similarity and it 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:

$$SR(v, v') = \frac{\gamma}{|in(v)| \cdot |in(v')|} \sum_{n \in in(v)} \sum_{n' \in in(v')} SR(n, n') \tag{5.4}$$

It has been shown in [32] that that $SR(v, v')$ is equivalent to the expected length (i.e., number of hops) of the first meeting of two coupled backward random walks, one starting from v and one starting from v'. This gives rise to a highly efficient computation proposed by Fogaras et al. [32]:

1. Compute (standard) random walks of bounded lengths with each node as a starting point.

2. Organize the reached nodes and their meeting distances in a *fingerprint tree (FPT)*. Given a set of coupled random walks, a fingerprint tree stores only the first meeting time for each pair of walks without capturing their original paths.

3. Repeat the random walks with different random choices creating m i.i.d. (independently and 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

Definition 5.4. We have extended the notion of SimRank by allowing weights for all edges, leading to the definition of *extended SimRank*:

$$w\text{-}SR(v,v') = \gamma \sum_{n \in In(v)} \sum_{n' \in In(v')} w(n \to v) \cdot w(n' \to v') \cdot SR(n,n') \qquad (5.5)$$

This extension is still equivalent to the expected meeting distance for the corresponding weighted (i.e., non-uniform) backward coupled random walks.

Another deviation from the standard SimRank is that we allow conceptual self-loops, introducing a bias towards reaching local-neighborhood nodes (i.e., penalizing long-distance walks). The weights of the edges adjacent to node v are uniformly distributed over the specific edge types:

1. *i-i* edges in the same language;

2. *i-c* edges in the same language;

3. *i-i* edges or *c-c* edges across languages.

The notion of edge weights can be further refined to accommodate the size of categories or the article length to bias the choice among link destinations. In the current implementation, LAIKA uses only edge-type-specific weights.

Novelty Ranking

This metric is specific to the knowledge repository scenario, specifically targeting recommendation of new categories to an article. The rationale behind this method is to propose a relevant but unexpected (novel) category for an article i relative to the already known categories of i.

To illustrate what we mean by novelty, consider the page about the former Prime Minister of France *Jean-Marc Ayrault*. Assume, we obtain two candidate categories: *Members of the French Socialist Party* and *Teachers*. The latter is preferred, since it is relevant but the most distant candidate to the already known categories for *Jean-Marc Ayrault*.

We implemented this idea by extending SimRank and denoting it as SR^*. We consider $n+1$ random walks starting at nodes $c_1, c_2, \ldots, c_n \in C(KR_n)$, the known categories of a source article i, and a recommendation candidate category c'. We compute the expected length until all $n+1$ walks meet, using the fingerprint trees.

Definition 5.5. Based on the expected meeting time of random walks starting at $c_1, c_2, \ldots, c_n \in C(KR_n)$, we define *novelty* of category c' with respect to the already known categories $c_1, c_2, \ldots, c_n \in C(KR_n)$ as:

$$Novelty(c') = 1 - SR^*(c_1, c_2, \ldots, c_n, c') \tag{5.6}$$

The computation determines the maximum distance over the meeting points and then averages over all fingerprint trees.

5.7 Experimental Evaluation

5.7.1 Experimental Setup

We 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 consider a pair of pages in two Wikipedia editions to be equivalent if the pages are connected with a bi-directional interlanguage link. Tables 5.2 and 5.3 summarize the resulting datasets and their cross-linkage.

For the SimRank-based methods we compute 400 fingerprints (i.i.d. random walks) of maximum length 100. In total, for all 5.5 Million nodes in our graph,

Wikipedia Edition	**# articles**	**# categories**	**# links**
German	2 338 795	139 844	45 531 135
French	2 408 097	199 708	42 022 704
Hungarian	339 041	34 653	6 273 337

Table 5.2: **Sizes of Wikipedia editions in March 2012.**

Inter-language Link Type	**German-French**	**German-Hungarian**	**French-Hungarian**
article-article	482 196	108 949	119 559
category-category	22 175	4 840	5 387

Table 5.3: **Existing alignments (inter-language links) between a pair of Wikipedia editions in March 2012.**

this precomputation took ca. 3 hours on a Linux server with 32 CPUs (3.2 GHz) and 503 Gb of RAM.

5.7.2 Ground Truth Construction

For each of the three link-prediction types, we generated a set of test cases with well-defined ground truth. We randomly removed 10% of the existing inter-language links, the article-category links (in the target knowledge repository KB_n), and article-article links (in the target knowledge repository KB_n), respectively. Then we predicted and ranked missing links, and compare the ranked results of the different recommendation methods against the originally existing links. In total, for all 6 language pairs the ground truth comprises 13,000 links for alignment curation task, 914,000 links for article-category recommendation, and 8.5 million links between related pages.

5.7.3 Measures

To assess quality of LAIKA's recommendations, we compute several standard measures.

Let $Cand(v)$ be the ranked list of targets for a node v, as introduced in Section 5.4. $|Cand(v)|$ is the length of the ranked list and $rel(i)$ be the relevance judgement of the i-th element in this ranked list.

The measures we report on are:

1. **Mean Reciprical Rank (MRR)**: the reciprocal of the highest rank at which a correct result appears. The formula for computing MRR was given earlier, in Section 3.9.3.

2. **Recall:** the fraction of ground-truth links recommended by a method. Let $Cand'(v)$ be the set of ground-truth targets for a node v. The following formula defines recall of recommendations $Cand(v)$:

$$recall(Cand(v)) = \frac{|\{i : rel(i) = correct\}|}{|Cand'(v)|} \tag{5.7}$$

3. **Precision:** the fraction of correct links among the recommended ones:

$$prec(Cand(v)) = \frac{|\{i : rel(i) = correct\}|}{|Cand(v)|} \tag{5.8}$$

4. **Precision@10:** the precision at the top-10 ranks only:

$$prec@10(Cand(v)) = \frac{|\{i : rel(i) = correct \ \& \ i <= 10\}|}{\min(|Cand(v)|, 10)} \tag{5.9}$$

5. **Normalized Discounted Cumulative Gain (nDCG):** the accumulated precision over all ranks, with ranks weighted in a geometrically decreasing manner (a standard measure for rankings in information retrieval).

   The discounted cumulative gain (DCG) is defined as:

$$DCG(Cand(v)) = \sum_{i=1}^{|Cand(v)|} \frac{rel(i)}{\log_2(i+1)} \tag{5.10}$$

   The ideal cumulative gain is achieved when a method returns candidates in the order of their relevance starting from most relevant ones:

$$IDCG(Cand(v)) = DCG(relCand(v)) \tag{5.11}$$

   where $relCand(v)$ denotes a list of only correct candidates ranked by their relevance.

Method	**MRR**	**nDCG**	**Recall**	**Precision**	**Precision@10**
		Alignment Curation: Predicting Missing Alignments			
Jaccard	0.539	0.764	0.630	0.214	0.227
SimRank	0.518	0.645	0.630	0.214	0.219
Voting	**0.712**	**0.850**	0.630	0.214	**0.230**
		Knowledge Repository Curation: Predicting New Categories			
Jaccard	0.734	0.857	0.367	0.291	0.291
SimRank	0.757	0.883	0.367	0.291	0.291
Novelty	**0.762**	**0.910**	0.367	0.291	0.291
		Knowledge Repository Curation: Predicting Related Articles			
Jaccard	**0.787**	**0.539**	0.165	0.062	**0.068**
SimRank	0.781	0.518	0.165	0.062	0.065

Table 5.4: **Results for three prediction tasks for alignment curation and knowledge repository curation.**

The normalized DCG is a ratio between the DCG of the recommended list of candidates to the DCG of the ideal recommendation:

$$nDCG(Cand(v)) = \frac{DCG(Cand(v))}{IDCG(Cand(v))} \tag{5.12}$$

The measures were computed for each query nodes v in the sample set and the averaged results are reported.

5.7.4 Results

Table 5.4 summarizes performance of LAIKA with respect to the three curation scenarios. The highest values in each column are marked with bold font. In this subsection we consider how scoring methods performed in each use case. Anecdotal examples in Table 5.5 illustrate performance of LAIKA per curation scenario.

Alignment Curation: Recommending Missing Alignments

For this type of curation task, there is only one correct result in the ground truth since Wikipedia allows linking a category to at most one equivalent category

Curation Task	Source	Targets
Alignment Curation: Predicting Missing Alignments	Seltsame Materie (en.: strange matter)	Csillagászati alapfogalmak (en.: basic astronomical concepts)
Knowledge Repository Curation: Predicting New Categories	Kosmische Strahlung (en.: cosmic ray)	Astrophysik (en.: astrophysics) Elektromagnetisches Spektrum (en.: electromagnetic spectrum) Teilchenphysik (en.: particle physics)
Knowledge Repository Curation: Predicting Related Articles	Kosmische Strahlung (en: cosmic ray)	Pierre Auger Arthur Holly Compton Charles Thomson Rees Wilson Teilchenphysik (en.: particle physics) Elementarteilchen (en.: elementary particles) Strahlung (en.: radiation) Partikel (en.: particle)

Table 5.5: **Anecdotal examples of found alignments and links.**

in a parallel language edition. Therefore, we concentrate on MRR and NDGC values as quality indicators.

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 counterpart was found on rank 1 or 2; in other words, nearly perfect predictions. The recall value of 0.63 indicates that in more than half of the instances, we found the correct alignment (not necessarily always in the top ranks, though).

From the example in Table 5.5, German category *Seltsame Materie* with only 9 pages therein still could get an aligned category in Hungarian Wikipedia edition. Note that all of these are small categories in the long tail. For example, the Hungarian Wikipedia does not contain an article on strange matter at all. The recommendations produced by LAIKA are not obvious, and the high accuracy of our methods is remarkable.

Knowledge Repository Curation: Predicting Related Categories for an Article

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 5.4 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 extended Novelty method is excelling.

The recall and precision are the same for all methods, as they worked on the same candidate sets ($Cand(v)$ as introduced in Section 5.4), solely ranking them differently. The recall of ca. 36% indicates 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. For the Novelty method, this effect also led to many ties in the scoring (of seemingly perfect score 1), which were broken at random. The SimRank-based methods could potentially overcome this current limitation in recall, by increasing the number of precomputed FPTs (i.i.d. walks).

Knowledge Repository Curation: Predicting Related Articles for an Article

As in the previous scenario, for a source article there are multiple related articles, but their number is usually much higher than in the category recommendation case. 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). Take sample recommendation given in Table 5.5 as example. By limiting the search space to people pages, the recommendations are *Pierre Auger*, *Arthur Holly Compton*, and *Charles Thomson Rees Wilson*. The recall numbers are also smaller than for the related categories recommendation case. Here, the much larger candidate sets aggravated the problem of zero overlap or non-meeting random walks.

5.8 Related Work

In this section we discuss the connections between LAIKA and the previous approaches. We categorize them into the three following groups.

Link prediction in Wikipedia. Chernov et al. [18] find semantic relations between Wikipedia categories based on interlinkage of pages belonging to categories, all within a single Wikipedia edition. Wu et al. [108] address the problem of automatically generating links between Wikipedia articles, using NLP and learning techniques. Both works can be considered as an approach to knowledge repository curation. LAIKA relies not only on the internal link structure, but also profits from the aligned link graph to detect missing connections for articles. In addition, LAIKA is also able to address the alignment curation problem targeting missing interlanguage links between two Wikipedia editions. Spiegel et al. [90] address 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 does not discriminate long-tail articles and small categories with sparse linkage.

Multilingual Wikipedia. De Melo et al. [20] use 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. This can also be considered as a task of alignment curation. However, LAIKA focuses on detecting missing alignments - non-existing inter-language links. Sorg et al. [88] consider a pair of Wikipedia editions to detect missing cross-language links between articles. The solution involves SVM classification, using a variety of link and content features. Wang et al. [106] pursue 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 content features. In contrast to these learning approaches, LAIKA is very efficient and is able to predict all kinds of missing links, most notably, interwiki links for small categories and categories for long-tail articles. Nguyen et al. [74] automatically match infobox schemas across multiple languages in Wikipedia. This data-integration task is tailored to detecting equivalences over infobox attributes and is very different from our mission of finding missing links for alignment and knowledge base curation.

Large-scale similarity computation. Lizorkin et al. [60] propose 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. Fogaras et al. [32] 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. Our scoring models have strong parallels with Milne et al. [67] and Turdakov et al. [101]. In both works a weighting scheme for edges in a Wikipedia graph is used, either based on the node degree or the type of Wikipedia pages an edge connects. However, using SimRank enables computing similarity based on the global link structure, rather than on solely considering node's direct neighbourhood.

5.9 Summary

This chapter presented the LAIKA system for dealing with alignment curation and knowledge base curation for highly dynamic and large link graphs. The main goal of LAIKA is to automatically generate recommendations of missing inter- and intra-lingual links to Wikipedia authors of different languages. We cast curation problems into link recommendation tasks and our experiments show that there is great potential for helping authors in dealing with long-tail entities and events.

Chapter 6

Conclusion

6.1 Summary of Results

Applications that span multiple knowledge repositories of different origins are of increasing interest. The heart of the intelligent systems of this kind is the ability to reconcile dynamic knowledge repositories with low affinity in terminology and structure. Within this dissertation we presented three main contributions to narrow the terminological and the structural gaps, as well as to maintain alignments between dynamic knowledge repositories.

Our first contribution is ACROSS, a system for reducing terminological heterogeneity. It has two major building blocks – semantification and constraint-aware reasoning. In the first step, ACROSS annotates the categories of the input knowledge repositories with instances and classes of a reference taxonomy (e.g., Wikipedia). This accounts for language variety, including synonymy and multilingualism. The reasoning step aims at filtering out noisy mappings, resolve ambiguities and produce concise and clean alignments, ready for human consumption. Our results demonstrated that ACROSS is able to find alignments, even for the categories which are deep in the input taxonomies, with high accuracy. In the experiment with health taxonomies, ACROSS reached 1.0 MRR. Injecting a small number of seed alignments had a drastic impact on the reasoning run time. Our experiments with the categorizations of software products, the run time was reduced by a factor of 45.

The second contribution is the LILIANA and SESAME systems. Their goal is to tackle the problem of aligning knowledge repositories given their low structural affinity. Both systems rely on a set of shared categories or instances between the input KRs to bootstrap the alignment procedure. To compensate for small coverage, the KR's link structures are harnessed to extend the alignment.

This allows for a larger discrepancy in the structures of the input repositories. Experimental evaluation of SESAME performance showed that contextualization via link structures is more robust than a plain keyword search in case of complex queries. In our experiments, SESAME reached 0.69 MRR compared to the performance of Bing and Google (0.41 and 0.24 respectively). In addition, we have presented the implementation of user interfaces of LILIANA and SESAME and gave examples of use-case scenarios.

Our final contribution is the LAIKA system targeting alignment curation for dynamic knowledge repositories. When aligning knowledge repositories, their dynamics becomes an obstacle for ensuring alignment completeness. Within our LAIKA system, we implemented an algorithm that accelerates detecting missing alignments. In addition, we investigated two scenarios of repository maintenance to transfer missing knowledge from the contextualized counterpart. For ranking candidate alignments and missing links, we studied several approaches based either on the local or on the global link structure. Our experimental evaluation showed that LAIKA is able to recommend missing inter- and intra-repository links with a high accuracy, reaching 0.78 MRR for the alignment curation task.

6.2 Future Directions

This dissertation has made contributions towards the integration of heterogeneous knowledge repositories. However, there are several research directions that are yet unaddressed. In what follows we give an overview of possible opportunities for future work.

Quantifying Heterogeneity

Throughout this dissertation we assumed the same degree of heterogeneity for all the objects of a KR. In reality, there are often subtrees of subgraphs common to several KRs, whereas some parts of the KRs are highly dissimilar and need more careful treatment.

For illustration, consider the top level categories of the book departments of Amazon.com and Amazon.de depicted in Figure 6.1. The majority of the categories is shared. In contrast, subtrees rooted at *Religion & Spirituality* are structured differently (cf. Figure 6.2). This calls for developing a systematic approach to quantifying heterogeneity.

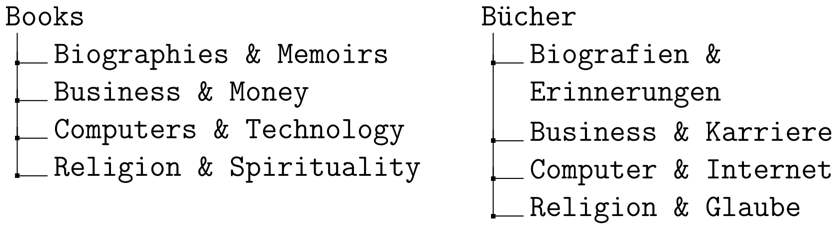


Figure 6.1: **Examples of coherent subtrees in Amazon's book departments (Germany- and US-based).**

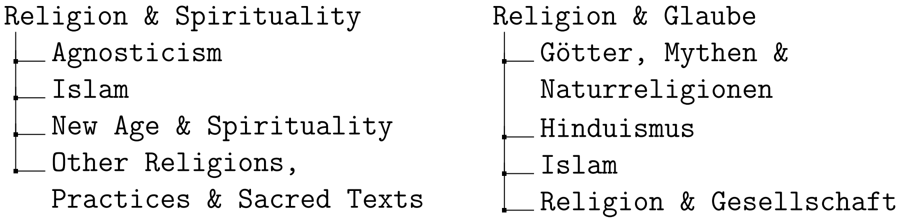


Figure 6.2: **Examples of highly dissimilar subtrees in Amazon's book departments (Germany- and US-based).**

Alignment Curation at Scale

Data analytics at scale has been gaining an increasing interest over the past years. It has been supported by the development of distributed frameworks as well as the availability of cloud platforms. These infrastructures together with the growing size of the Web become of high relevance to the data integration and ontology alignment research. However, classical ontology mapping solutions are limited to inputs of modest size and treat input sources as static snapshots. Scaling alignment methods to knowledge repositories of large size and high dynamics still remains an open problem.

We have addressed the problem of large input sizes in the ACROSS system. In order to reduce the run time of the reasoning procedure, we proposed to use a set of anchor alignments. However, interlinking highly dynamic KRs is a challenging task for ACROSS. To make an alignment decision for a category, ACROSS needs to construct candidate alignments for the entire taxonomy and perform joint reasoning.

In contrast to ACROSS, the LILIANA, the LAIKA and the SESAME systems explore only the local link structure of a node. This make them suitable for interlinking large and dynamic knowledge repositories. The open challenge for these systems is, however, to consider additional constraints while constructing an alignment.

Uncovering Fake Concepts

News verification and credibility assessment has been gaining increasing attention over the past years. Some approaches have been proposed by Jin et al. [50] and Popat et al. [81, 82]. These are, however, grounded in natural language features and do not utilize composite knowledge available in LOD-like sources. Future research might therefore draw more strength from contextualization of news articles by mapping them onto a set of interlinked knowledge repositories. Possible directions include, but are not limited to, learning recurring patterns of fake news and identifying concepts with high heterogeneity across sources.

Bibliography

[1] R. Agrawal and R. Srikant. On Integrating Catalogs. In *Proceedings of the 10th International Conference on World Wide Web*, pages 603–612, 2001.

[2] A. Alawini, D. Maier, K. Tufte, B. Howe, and R. Nandikur. Towards Automated Prediction of Relationships Among Scientific Datasets. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, pages 35:1–35:5, 2015.

[3] Z. Aleksovski, M. Klein, W. Ten Kate, and F. Van Harmelen. Matching Unstructured Vocabularies Using a Background Ontology. *Managing Knowledge in a World of Networks*, pages 182–197, 2006.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web*, pages 722–735, 2007.

[5] K. Balog, M. Bron, and M. De Rijke. Query Modeling for Entity Search Based on Terms, Categories, and Examples. *ACM Transactions on Information Systems (TOIS)*, 29(4):22, 2011.

[6] S. Banerjee and T. Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145, 2002.

[7] S. Banerjee and T. Pedersen. Extended Gloss Overlaps As a Measure of Semantic Relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 805–810, 2003.

[8] C. S. Bhagavatula, T. Noraset, and D. Downey. TabEL: Entity Linking in Web Tables. In *Proceedings of the International Semantic Web Conference*, pages 425–441, 2015.

[9] O. Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

[10] N. Boldyrev, M. Spaniol, J. Strötgen, and G. Weikum. SESAME: European Statistics Explored via Semantic Alignment onto Wikipedia. In *Proceedings of the 26th ACM International Conference on World Wide Web Companion*, pages 177–181, 2017.

[11] N. Boldyrev, M. Spaniol, and G. Weikum. ACROSS: A Framework for Multi-cultural Interlinking of Web Taxonomies. In *Proceedings of the 8th ACM Conference on Web Science*, pages 127–136, 2016.

[12] N. Boldyrev, M. Spaniol, and G. Weikum. Multi-Cultural Interlinking of Web Taxonomies with ACROSS. *The Journal of Web Science*, under review.

[13] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250, 2008.

[14] N. Bronson, Z. Amsden, G. Cabrera, P. Chakka, P. Dimov, H. Ding, J. Ferris, A. Giardullo, S. Kulkarni, H. C. Li, et al. TAO: Facebook's Distributed Data Store for the Social Graph. In *USENIX Annual Technical Conference*, pages 49–60, 2013.

[15] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the 24th Conference on Artificial Intelligence*, pages 1306–1313, 2010.

[16] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684, 2011.

[17] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[18] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting Semantics Relationships between Wikipedia Categories. In *Proceedings of the First Workshop on Semantic Wikis – From Wiki to Semantics*, 2006.

[19] M. Y. Dahab, H. A. Hassan, and A. Rafea. TextOntoEx: Automatic Ontology Construction from Natural English Text. *Expert Systems with Applications*, 34(2):1474–1480, 2008.

[20] G. de Melo and G. Weikum. Untangling the Cross-lingual Link Structure of Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 844–853, 2010.

[21] R. Dhamankar, Y. Lee, A. Doan, A. Halevy, and P. Domingos. iMAP: Discovering Complex Semantic Matches Between Database Schemas. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, pages 383–394, 2004.

[22] H.-H. Do and E. Rahm. COMA: a System for Flexible Combination of Schema Matching Approaches. In *Proceedings of the 28th International Conference on Very Large Data Bases*, pages 610–621, 2002.

[23] A. H. Doan and A. Y. Halevy. Semantic integration research in the database community: A brief survey. *AI Magazine*, 26(1):83–94, 2005.

[24] D. Dou and P. LePendu. Ontology-based Integration for Relational Databases. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, pages 461–466, 2006.

[25] W. Drabent, T. Eiter, G. Ianni, T. Krennwallner, T. Lukasiewicz, and J. Małuszyński. *Hybrid Reasoning with Rules and Ontologies*, pages 1–49. Springer-Verlag, 2009.

[26] W. Drabent and J. Małuszyński. Well-Founded Semantics for Hybrid Rules. *Web Reasoning and Rule Systems*, pages 1–15, 2007.

[27] S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum. Language-model-based Ranking for Queries on RDF-graphs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 977–986, 2009.

[28] P. Ernst, C. Meng, A. Siu, and G. Weikum. KnowLife: a Knowledge Graph for Health and Life Sciences. In *Proceedings of the 30th IEEE International Conference on Data Engineering*, pages 1254–1257, 2014.

[29] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale Information Extraction in KnowItAll: (Preliminary Results). In *Proceedings of the 13th International Conference on World Wide Web*, pages 100–110, 2004.

[30] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg, 2nd edition, 2013.

[31] J. L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York, 1973.

[32] D. Fogaras and B. Rácz. Scaling link-based similarity search. In *Proceedings of the 14th ACM International Conference on World Wide Web*, pages 641–650, 2005.

[33] A. Fokoue, O. Hassanzadeh, M. Sadoghi, and P. Zhang. Predicting Drug-Drug Interactions Through Similarity-Based Link Prediction Over Web Data. In *Proceedings of the 25th ACM International Conference on World Wide Web Companion*, pages 175–178, 2016.

[34] B. Fu, R. Brennan, and D. O'Sullivan. Cross-Lingual Ontology Mapping – An Investigation of the Impact of Machine Translation. In *Proceedings of the 4th Asian Semantic Web Conference*, pages 1–15, 2009.

[35] P. Gamallo, C. Gasperin, A. Agustini, and G. Lopes. Using Syntactic Contexts for Measuring Word Similarity. In *Proceedings of the 4th International Conference on Text, Speech and Dialogue*, 2001.

[36] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-Match: an Algorithm and an Implementation of Semantic Matching. In *Proceedings of the European Semantic Web Symposium*, pages 61–75, 2004.

[37] J. Göbölös-Szabó, N. Prytkova, M. Spaniol, and G. Weikum. Cross-Lingual Data Quality for Knowledge Base Acceleration across Wikipedia Editions. In *Proceedings of the 10th International Workshop on Quality in Databases*, pages 1–7, 2012.

[38] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In *Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining*, pages 1019–1028, 2010.

[39] J. Gracia and E. Mena. Web-based Measure of Semantic Relatedness. In *Proceedings of the International Conference on Web Information Systems Engineering*, pages 136–150, 2008.

[40] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae. Challenges for the Multilingual Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71, 2012.

[41] A. Y. Halevy, N. Ashish, D. Bitton, M. Carey, D. Draper, J. Pollock, A. Rosenthal, and V. Sikka. Enterprise Information Integration: Successes, Challenges and Controversies. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pages 778–787, 2005.

[42] S. Hertling and H. Paulheim. WikiMatch: Using Wikipedia for Ontology Matching. In *Proceedings of the 7th International Conference on Ontology Matching*, pages 37–48, 2012.

[43] J. Hoffart, D. Milchevski, and G. Weikum. STICS: Searching with Strings, Things, and Cats. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1247–1248. ACM, 2014.

[44] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792, 2011.

[45] W. Hu, Y. Qu, and G. Cheng. Matching Large Ontologies: A Divide-and-Conquer Approach. *Data and Knowledge Engineering*, 67(1):140–160, 2008.

[46] R. Ichise, H. Takeda, and S. Honiden. Integrating Multiple Internet Directories by Instance-based Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 3, pages 22–28, 2003.

[47] P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh. Ontology Alignment for Linked Open Data. In *Proceedings of the International Semantic Web Conference*, pages 402–417, 2010.

[48] G. Jeh and J. Widom. SimRank: A Measure of Structural-context Similarity. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543, 2002.

[49] E. Jiménez-Ruiz and B. Cuenca Grau. LogMap: Logic-Based and Scalable Ontology Matching. In *Proceedings of the 10th International Semantic Web Conference*, pages 273–288, 2011.

[50] Z. Jin, J. Cao, Y. Zhang, and J. Luo. News Verification by Exploiting Conflicting Social Viewpoints in Microblogs. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, pages 2972–2978, 2016.

[51] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, 2009.

[52] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.

[53] M. Lenzerini. Data Integration: A Theoretical Perspective. In *Proceedings of the 21st ACM Symposium on Principles of Database Systems*, pages 233–246, 2002.

[54] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506, 2009.

[55] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[56] W.-S. Li and C. Clifton. SEMINT: A Tool for Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Networks. *Data and Knowledge Engineering*, 33(1):49–84, 2000.

[57] X. Li, W. Meng, and C. Yu. T-Verifier: Verifying Truthfulness of Fact Statements. In *Proceedings of the 27th IEEE International Conference on Data Engineering*, pages 63–74, 2011.

[58] G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and Searching Web Tables Using Entities, Types and Relationships. *Proceedings of the VLDB Endowment*, 3(1-2):1338–1347, 2010.

[59] F. Lin and W. W. Cohen. Semi-Supervised Classification of Network Data Using Very Few Labels. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 192–199, 2010.

[60] D. Lizorkin, P. Velikhov, M. Grinev, and D. Turdakov. Accuracy Estimate and Optimization Techniques for SimRank Computation. *Proceedings of the VLDB Endowment*, 1(1):422–433, 2008.

[61] W. Lu, J. Janssen, E. Milios, and N. Japkowicz. Node Similarity in Networked Information Spaces. In *Proceedings of the Conference of the Centre for Advanced Studies on Collaborative Research*, page 11, 2001.

[62] J. Madhavan, P. A. Bernstein, A. Doan, and A. Halevy. Corpus-based Schema Matching. In *Proceedings of the 21st International Conference on Data Engineering*, pages 57–68, 2005.

[63] J. Madhavan, P. A. Bernstein, and E. Rahm. Generic Schema Matching with Cupid. In *Proceedings of the 27th International Conference on Very Large Data Bases*, volume 1, pages 49–58, 2001.

[64] C. Meilicke. Alignment Incoherence in Ontology Matching. PhD Thesis, Universität Mannheim, 2011.

[65] S. Metzler, S. Günnemann, and P. Miettinen. Hyperbolae are no Hyperbole: Modelling Communities that are not Cliques. In *Proceedings of the 16th IEEE International Conference on Data Mining*, pages 330–339, 2016.

[66] G. A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[67] D. Milne and I. H. Witten. An Effective, Low-Cost Measure of SemanticRelatedness Obtained from Wikipedia Links. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence*, pages 25–30, 2008.

[68] P. Mirza, S. Razniewski, and W. Nutt. Expanding Wikidata's Parenthood Information by 178%, or How To Mine Relation Cardinality Information. In *Proceedings of the 15th International Semantic Web Conference Posters & Demonstrations Track*, 2016.

[69] A. Mishra, D. Milchevski, and K. Berberich. Linking Wikipedia Events to Past News. In *Proceedings of the SIGIR ACM 2014 Workshop on Temporal, Social and Spatially-aware Information Access*, pages 1–4, 2014.

[70] N. Nakashole, T. Tylenda, and G. Weikum. Fine-grained Semantic Typing of Emerging Entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1488–1497, 2013.

[71] F. Nanni, S. P. Ponzetto, and L. Dietz. Entity Relatedness for Retrospective Analyses of Global Events. In *Proceedings of the ACM Workshop on Natural Language Processing and Computational Social Science*, 2016.

[72] D. Nguyen, M. Theobald, and G. Weikum. J-NERD: Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features. *Transactions of the Association for Computational Linguistics*, 4:215–229, 2016.

[73] H.-V. Nguyen, P. Mandros, and J. Vreeken. Universal Dependency Analysis. In *Proceedings of the SIAM International Conference on Data Mining*, pages 792–800, 2016.

[74] T. Nguyen, V. Moreira, H. Nguyen, H. Nguyen, and J. Freire. Multilingual Schema Matching for Wikipedia Infoboxes. *Proceedings of the VLDB Endowment*, 5(2):133–144, 2011.

[75] Z. Nie, J.-R. Wen, and W.-Y. Ma. Statistical Entity Extraction from the Web. *Proceedings of the IEEE*, 100(9):2675–2687, 2012.

[76] N. F. Noy and M. A. Musen. Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence*, 2000.

[77] N. F. Noy and M. A. Musen. Anchor-PROMPT: Using Non-local Context for Semantic Matching. In *Proceedings of the Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence*, pages 63–70, 2001.

[78] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

[79] P. Papadimitriou, P. Tsaparas, A. Fuxman, and L. Getoor. TACI: Taxonomy-Aware Catalog Integration. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1643–1655, 2013.

[80] H. Paulheim. WeSeE-Match Results for OEAI 2012. In *Proceedings of the 7th International Conference on Ontology Matching*, pages 213–219, 2012.

[81] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. Credibility Assessment of Textual Claims on the Web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2173–2178, 2016.

[82] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th ACM International Conference on World Wide Web Companion*, pages 1003–1012, 2017.

[83] N. Prytkova, M. Spaniol, and G. Weikum. Aligning Multi-Cultural Knowledge Taxonomies by Combinatorial Optimization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 93–94, 2015.

[84] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum. YAGO: a Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In *Proceedings of the International Semantic Web Conference*, pages 177–185, 2016.

[85] P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[86] M. Sabou, M. d'Aquin, and E. Motta. *Exploring the Semantic Web as Background Knowledge for Ontology Matching*, pages 156–190. Springer Berlin Heidelberg, 2008.

[87] A. Singhal. Introducing the Knowledge Graph: Things, Not String. Official Blog of Google, 2012.

[88] 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*, pages 49–54, 2008.

[89] M. Spaniol, N. Prytkova, and G. Weikum. Knowledge Linking for Online Statistics. In *Proceedings of the 59th World Statistics Congress*, 2013.

[90] S. Spiegel, J. Clausen, S. Albayrak, and J. Kunegis. Link Prediction on Evolving Data Using Tensor Factorization. In *Proceedings of the 15th International Conference on New Frontiers in Applied Data Mining*, pages 100–110, 2012.

[91] A. Spitz and M. Gertz. Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 503–512, 2016.

[92] D. Spohr, L. Hollink, and P. Cimiano. A Machine Learning Approach to Multilingual and Cross-Lingual Ontology Matching. In *Proceedings of the 10th International Semantic Web Conference*, pages 665–680, 2011.

[93] S. Staab and R. Studer. *Handbook on ontologies*. Springer Science & Business Media, 2013.

[94] J. Strötgen and M. Gertz. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th ACL International Workshop on Semantic Evaluation*, pages 321–324, 2010.

[95] R. Studer, V. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1):161 – 197, 1998.

[96] F. M. Suchanek, S. Abiteboul, and P. Senellart. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. *Proceedings of the VLDB Endowment*, 5(3):157–168, 2011.

[97] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: a Core of Semantic Knowledge. In *Proceedings of the 16th ACM International Conference on World Wide Web*, pages 697–706, 2007.

[98] B. Taskar, P. Abbeel, M.-F. Wong, and D. Koller. Label and Link Prediction in Relational Data. In *Proceedings of the Workshop on Learning Statistical Models from Relational Data*, 2003.

[99] A. Tian, M. Kejriwal, and D. P. Miranker. Schema Matching over Relations, Attributes, and Data Values. In *Proceedings of the 26th ACM International Conference on Scientific and Statistical Database Management*, pages 28:1–28:12, 2014.

[100] Y. Tian, R. C. Mceachin, C. Santos, D. J. States, and J. M. Patel. SAGA: A Subgraph Matching Tool for Biological Graphs. *Bioinformatics*, 23(2):232–239, 2007.

[101] D. Turdakov and P. Velikhov. Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disambiguation. In *Proceedings of the 2008 Colloquium on Databases and Information Systems*, 2008.

[102] O. Udrea, L. Getoor, and R. J. Miller. Leveraging Data and Structure in Ontology Integration. In *Proceedings of the 2007 ACM International Conference on Management of Data*, pages 449–460, 2007.

[103] J. D. Ullman. Information Integration Using Logical Views. In *Proceedings of the 6th International Conference on Database Theory*, pages 19–40, 1997.

[104] M. Uschold and M. Gruninger. Ontologies and Semantics for Seamless Connectivity. *ACM SIGMod Record*, 33(4):58–64, 2004.

[105] Y. Velegrakis, R. J. Miller, L. Popa, and J. Mylopoulos. ToMAS: A System for Adapting Mappings while Schemas Evolve. In *Proceedings of the 20th IEEE International Conference on Data Engineering*, page 862, 2004.

[106] Z. Wang, J. Li, Z. Wang, and J. Tang. Cross-Lingual Knowledge Linking across Wiki Knowledge bases. In *Proceedings of the 21st ACM International Conference on World Wide Web*, pages 459–468, 2012.

[107] B. Wu and B. D. Davison. Identifying Link Farm Spam Pages. In *In Proceedings of the 14th ACM International Conference on World Wide Web, Special Interest Tracks and Posters*, pages 820–829, 2005.

[108] F. Wu and D. S. Weld. Autonomously Semantifying Wikipedia. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, pages 41–50, 2007.

[109] J. Yang and J. Leskovec. Community-Affiliation Graph Model for Overlapping Network Community Detection. In *In proceedings of the 12th IEEE International Conference on Data Mining*, pages 1170–1175, 2012.

[110] S. Zampelli, Y. Deville, and P. Dupont. Approximate Constrained Subgraph Matching. In *In Proceedings of the 11th International Conference on Principles and Practice of Constraint Programming*, pages 832–836, 2005.