Querying the Web with Statistical Machine Learning

Volker Tresp, Yi Huang, and Maximilian Nickel

Abstract The traditional means of extracting information from the Web are keyword-based search and browsing. The Semantic Web adds structured information (i.e., semantic annotations and references) supporting both activities. One of the most interesting recent developments is Linked Open Data (LOD) where information is presented in the form of facts —often originating from published domain-specific databases— that can be accessed both by a human and a machine via specific query endpoints. In this chapter we argue that machine learning provides a new way to query Web data, in particular LOD, by analyzing and exploiting statistical regularities. We discuss challenges when applying machine learning to the Web and discuss the particular learning approaches we have been pursuing in THESEUS. We discuss a number of applications, where the Web is queried via machine learning and describe several extensions to our approaches.

1 Introduction

The traditional means of extracting information from the Web are keyword-based search and browsing. In search, the user enters query terms and, if lucky, can read off the required information from the returned pages. In browsing, the user follows hyperlinks to gain deeper information on an issue. The Semantic Web adds structured information (i.e., semantic annotations and references) supporting both keyword-based search and browsing. One of the most interesting recent developments here is Linked Open Data (LOD) [3] where information is presented in form

Volker Tresp
Siemens AG, Otto-Hahn-Ring 6, 81739 Muenchen, e-mail: Volker.Tresp@Siemens.com

Yi Huang
Siemens AG, Otto-Hahn-Ring 6, 81739 Muenchen, e-mail: YiHuang@Siemens.com

Maximilian Nickel
Ludwig Maximilian University of Munich, e-mail: nickel@dbs.ifi.lmu.de

of facts —often originating from published domain-specific databases— that can be accessed both by a human and a machine via specific query endpoints. Thus, one can query for the "10 largest German companies whose CEOs were born in the US" or a list of "genes associated with a given disease". LOD does not just *reference* information, it *represents* information in form of simple subject-predicate-object triples. With this novel representation of information, new opportunities for accessing information emerge that explore and exploit regularities in the represented data. In recent years mostly deterministic regularities, which can be formulated as logical expressions, have been explored. Thus, deductive reasoning might conclude that an author born in Landshut would also be recognized as an author born in Bavaria. Deterministic regularities originate, for example, from natural laws (e.g., law of gravity), from human definitions and conventions (e.g.,"dogs are mammals"), from design (e.g.,"the car only starts when the key is turned"), and from human imposed laws and regulations (e.g.,"work begins at 9 am"). In addition to deterministic or close-to-deterministic regularities, the world also contains statistical regularities. One might debate if the world is inherently deterministic or probabilistic, but at the abstract level of representation, which is typically available for decision making, the world certainly appears uncertain.[1] Young males typically like action movies but whether young Jack will buy "Iron Man 2" might depend more on the availability of illegal downloads, the opinions of his peers and Jack's financial situation. A system recommending a movie to Jack must work with the available information, maybe a list of movies that Jack has bought before, and can only make statistical recommendations. Similarly, the interactions between genes and diseases might or might not be inherently deterministic in nature; at the level of current knowledge the relationships are only partially known.

Machine learning is a basis for extracting statistical patterns and in this chapter we will describe our work on statistical machine learning for the Web as pursued in THESEUS and in the EU FP7 project LarKC [5]. In this work we have proposed that statistical patterns extracted from the Web via machine learning should be integrated into queries [16]. Thus a search for diseases associated with a particular gene can be done in three ways: First, one can study the documents returned via keyword-based search. Second, one can obtain a list of diseases known to be, or suspected to be, associated via a structured query on LOD. Finally, one can use machine learning to extract diseases likely to be related to the gene based on disease and gene attributes and established gene-disease patterns. Note that machine learning depends on repeatable statistical patterns: thus machine learning cannot help to give you the first name of the wife of the US-president (a one-to-one relationship), but it can predict the probability of re-election, his income, the party of the vice president and the number of expected grand children.

In the next section we discuss some of the challenges encountered when applying machine learning to LOD. In Section 3 we motivate and describe our particular approaches. Section 4 describes a number of applications. One of them is BOTTARI,

[1] Although the world might be governed by scientific laws and logical constraints in general, at the level of abstraction that we and our applications have to function, the world partially appears to be governed by probabilities and statistical patterns.

the winning entry in the ISWC 2011 Semantic Web Challenge.[2] In Section 5 we describe extensions and future work. Section 6 contains our conclusions.

2 Challenges for Machine Learning

Machine learning is not only of interest to the Web, the Web also poses interesting research challenges to machine learning. First of all, Web data typically does not represent an i.i.d. (independent and identically distributed) statistical sample of some sort but might have been collected and published for any reason, often not following a particular systematics. For similar reasons, the data, in general, is incomplete, e.g., from the fact that a social network lists no friends of Jack one cannot conclude that Jack does not have any friends. In general, negation is very rare on Web data, thus one might find information that two persons are friends but rarely that two persons are not friends. This needs to be considered to avoid biased predictions. Another interesting property of Web data is that relationships between entities are often more informative than entity attributes, an effect exploited in collective learning: It might be easier to predict Jane's wealth from the wealth of her friends than from known properties of Jane. As in this example, nonlocal aggregated information is often informative for a prediction task and machine learning needs to take this into account. Sometimes, as in the examples mentioned in the introduction, relationships themselves are of interest, e.g., item preferences, friendships, relationships between genes and diseases. Since the number of potential relationships can be very large, the output of a learning system will often be a ranked list of candidate relationships, e.g., a ranked list of recommended items, instead of a single answer. As a particular feature of Web data, there is often textual information available that describes entities (e.g., Wikipedia articles), events (e.g., news stories) or topics (e.g., blogs) and this information can often be very useful for the machine learning task. Finally, a machine learning system has to be able to handle the large scale of the Web, its dynamical nature and its noisiness.

3 Predicting Facts with Factorization

In LOD, basic facts are represented as subject-predicate-object triples (s, p, o). In our work, we have been addressing the challenge of using machine learning to predict the likelihood of triples that are not explicitly given in the data. Since triples can describe class membership (Jane, rdf:type, Student), entity attributes (Jane, income, High) and relationships (Jane, likes, Jack), triple prediction is a quite general task. Equivalently, we might look at the LOD graph where the nodes represent the entities and a directed link represents an (s,p,o) triple, labeled by the predicate. In this view

[2] http://challenge.semanticWeb.org/2011/

the learning tasks consists of predicting the existence of labeled links not explicitly given in the graph.

Some of the most powerful learning approaches, effective for predicting links in a graph with properties as discussed in the last section, rely on a description of an entity in a latent space. What this means is that each entity is described by a number of features which might or might not have a real world meaning and which are abstracted from the information describing the entity. One example would be a cluster assignment, e.g., a student might belong to the cluster of "good students". Another abstract representation can be obtained via factorization approaches. Here an entity is described by degrees of agreement with certain factors. As an example, a student might agree with the factor "good students" with some degree, and with the factor "popular students" with some other degree. It turns out that latent factors are not simply helpful for understanding a domain but are also very effective in link prediction. In a factorization approach, the likelihood of the existence of a link is determined by the scalar product between the latent factors describing the associated entities. [3] [16, 6] describe the SUNS framework, which is the particular factorization approach developed in our work.

Factorization approaches provide high-quality predictions and are robust to the challenges described in Section 2. In particular they are highly scalable by exploiting the sparsity in the data and are suitable for making use of relationship information and for predicting relationships. Extensions towards deductive reasoning, for the inclusion of textual information, and for addressing the dynamical nature of data will be discussed in Section 5

4 Querying the Web with Machine Learning

In this section we describe some applications of machine learning to LOD. For more details, see the respective references.

---

[3] In particular, the probability that a relationship between two entities exists given the knowledge base KB is estimated as

$$\hat{P}((Jane, likes, Jack)|KB) = \sum_{i=1}^{L} f_i^{Jane} f_i^{likes,Jack} = f^{Jane^T} f^{likes,Jack}$$

where $(f_1^{Jane}, f_2^{Jane}, \ldots, f_L^{Jane})^T$ is the vector of L factors describing Jane, and $(f_1^{likes,Jack}, f_2^{likes,Jack}, \ldots, f_L^{likes,Jack})^T$ are the L factors describing Jack in his role as an object of the predicate "likes". $(.)^T$ denotes the transposed of a vector. There are a number of approaches for calculating the factors. In our work in the SUNS framework [16, 6], we have employed regularized factorization of the associated data matrices. In our 3-way tensor approach RESCAL [14], we estimate

$$\hat{P}((Jane, likes, Jack)|KB) = f^{Jane^T} R^{likes} f^{Jack}.$$

Each entity has a unique latent representation and the relation-type specific interaction is modeled by the $L \times L$ matrix R^{likes}.

4.1 Querying Social Networks

The experiments presented in this section are based on friend-of-a-friend (FOAF) data, which is part of LOD. The purpose of the FOAF project [4] is to create a Web of machine-readable pages describing people, their relationships, people's activities and their interests using W3C's RDF technology. The population is defined by the 32,062 persons in our FOAF subset. 14,425 features are formed by potential friends in the data. Furthermore, 781 attributes refer to general information about age, location, number of blog posts, attended school, online chat accounts and interests. The task is to predict potential friends of a person, i.e., *knows* statements, and the performance is evaluated using a test set of known friendships. In a comparison with competing methods, the factorization approach gave best performance in predicting new friendships [6]. The following SPARQL expression illustrates a query for LiveJournal[4] users who live in Munich and might want to be Trelena's friend:

```
PREFIX ya: http://blogs.yandex.ru/schema/foaf/
PREFIX foaf: http://xmlns.com/foaf/0.1/
PREFIX dc: http://purl.org/dc/elements/1.1/
SELECT DISTINCT ?person
WHERE {
  ?person ya:located ?city .
  ?person foaf:knows <http://trelana.livejournal.com/trelana>
  WITH PROB ?prob .
  FILTER REGEX(?city, Munich) .
}
ORDER BY DESC(?prob)
```

Listing 1 The query includes the predicted *knows* triples for Trelena and rates them by predicted probability.

The query contains an extended clause WITH PROB that returns the estimated probabilities of a friendship relationship for Trelena, modeled by the *knows* relation. Figure 1 presents a typical query response.

4.2 Querying Linked Life Data

Life science data forms a significant part of the LOD cloud. To a large extent, this data has been extracted from well-maintained databases such that this portion of LOD is of high quality. We applied our approach to an important problem in the life sciences, i.e., the prediction of gene-disease relationships, and demonstrated that we obtained competitive results to state-of-the-art solutions.

Disease genes are those genes involved in the causation of, or associated with, a particular disease. At this stage, more than 2500 disease genes have been discovered. Unfortunately, the relationship between genes and diseases is far from simple

[4] http://www.livejournal.com/

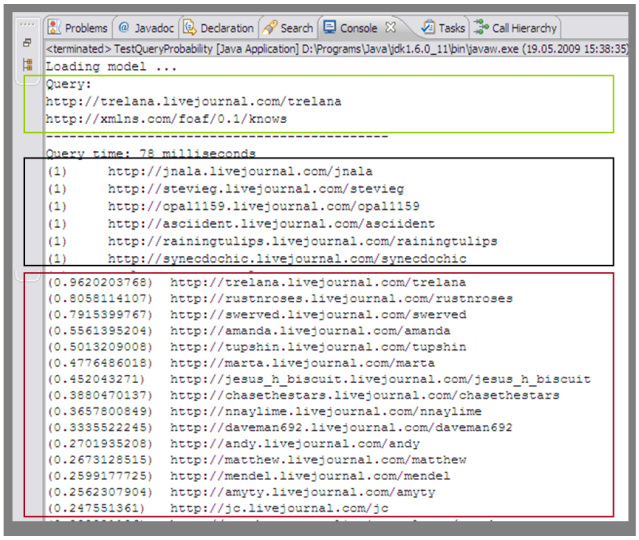

Fig. 1 Query: Who wants to be Trelena's friend. Her actual friends are predicted first with confidence values one. Then, interestingly, it is predicted that she should be her own friend, followed by a ranked list of predicted friends.

since most diseases are polygenic and exhibit different clinical phenotypes. High-throughput genome-wide studies like linkage analysis and gene expression profiling typically result in hundreds of potential candidate genes and it is still a challenge to identify the disease genes among them. One reason is that genes can often perform several functions and a mutational analysis of a particular gene reveals dozens of mutation sites that lead to different phenotype associations to diseases like cancer [12]. Analysis is further complicated when environmental and physiological factors come into play as well as by exogenous agents such as viruses and bacteria.

Despite this complexity, it is quite important to be able to rank genes in terms of their predicted relevance for a given disease. Such a ranking cannot only be a valuable tool for researchers but also has applications in medical diagnosis, prognosis, and a personalized treatment of diseases.

Gene properties differentiate disease genes and have been used as the bases for computational tools to prioritize disease gene candidates. All of the current approaches are based on the integration of different properties such as: gene function (disease genes are expected to share common functional properties), pathways (disease genes are most likely to share common pathways), gene expression (disease genes are expected to be co-expressed), gene regulation (genes within the same gene-regulation network are expected to affect similar diseases), sequence properties, and protein interaction (disease genes are often highly connected with other genes from the same disease). These attributes have been used in our approach as well. In addition, our approach exploits gene-disease interaction patterns. The solution was integrated into the THESEUS MEDICO use case. In 7 out of 12 experi-

ments the machine learning approach using the SUNS framework was superior to a leading competing approach based on a fuzzy-based similarity measure. See [8] for a detailed description of experimental results.

4.3 BOTTARI: Personalized and Location-based Recommendations

BOTTARI is an augmented reality application that permits the personalized and localized recommendation of points of interests (POIs) based on the temporally weighted opinions of the community. Opinions on POIs (here: restaurants) were extracted from Twitter[5] microposts with natural language processing and a background ontology. The task of the machine learning module was to rank different restaurants based on the extracted information and based on individual preference profiles of users. A successful evaluation of BOTTARI was carried out using a three year collection of tweets of about 319 restaurants located in the 2 km^2 district of Insadong, a popular tourist area of the South Korean city of Seoul (Figure 2).

The technological basis of BOTTARI is the highly scalable LarKC platform for the rapid prototyping and development of Semantic Web applications. BOTTARI is the winner of the 9th edition of the Semantic Web Challenge, co-located with the 2011 International Semantic Web Conference. BOTTARI is currently field tested in Korea by Saltlux [2].

4.4 More Examples

DBpedia [1] is part of LOD and contains structured information extracted from Wikipedia. At the time of the experiment, it describes more than 3.4 million concepts, including, e.g., 312,000 persons, 413,000 places and 94,000 music albums, DBpedia does not only serve as a "nucleus for the Web of data", but also holds great potential to be used in conjunction with machine learning approaches. DBpedia already provides a great value and is useful for accessing facts by, e.g., answering queries for the famous citizens and the most spectacular sights of a large number of cities. DBpedia data is increasingly getting interesting for machine learning purposes as new and richer relationships are added. In our experiments, we used a population consisting of all members of the German Bundestag to evaluate our approach [7]. The task was to predict the party membership based on age, state of birth, and keywords from the Wikipedia pages of the Bundestag members as input information. Most informative were the latter two sources, in particular also the state information, which can be explained by the peculiarities of German politics. Overall, we obtained a system that achieved high accuracy. Although there might not be

[5] https://twitter.com/

Fig. 2 A picture of Insadong district: the density of restaurants is very high.

any economic value in this experiment, it provides first insight into how DBpedia can be queried in the future using machine learning.

The SUNS approach was used in a prototype for analyzing the usage service patterns and for recommending services in a Web service platform as part of the THESEUS TEXO use case and in a prototype application for airline claim managements, presented at the CeBIT 2011.

5 Extensions

The factorization approaches briefly described in Section 3 are surprisingly general and powerful. In this section we describe some important extensions. As mentioned, we use machine learning to predict triples, resp. labeled links, based on statistical patterns in the data whereas deductive reasoning derives triples using facts and logical expressions, such as rules. Machine learning can easily benefit from deductive reasoning by including the derived triples in training and prediction.

Deductive reasoning can be helpful for aggregation as well. In many applications, information that is not local might become relevant for machine learning. As an example, in the DBpedia experiment party membership is more easily predicted

from a politician's state of birth than from a politician's city of birth, however only the latter information is explicitly stated in DBpedia. The state can be derived by using geo-reasoning from the city prior to learning. Materialization of knowledge by making implicit knowledge explicit via computing the inductive closure is a highly scalable approach to reasoning [13].

Important sources of information are often documents describing the involved entities or relations between entities as shown in the DBpedia experiments in Section 4.4 and textual information can support triple prediction in general. The combination of information extraction from text, deductive reasoning and machine learning to improve triple prediction is described in a probabilistic extension of the factorization approach in [9, 11].

Another interesting aspect concerns sequence and temporal information. Often a series like "Star Wars" will be watched in order. Similarly, medical procedures are given in a sensible sequential order. In [17] an extension to the factorization model is described that can model both sequential information and absolute time.

In the SUNS approach, triples were mapped to one or several matrices and the latent factors were calculated via a factorization of these matrices. It might be argued that a more natural representation for LOD's triple structure is given by a three-way tensor. Whereas in a matrix an element is addressed by two indices, in a 3-way tensor an entry is addressed by three indices. Reference [14] describes an approach where all entities of a domain are mapped to two modes of a 3-way tensor and the predicates are mapped to the third mode. In this tensor, an element equal to one indicates that the corresponding (s, p, o) triple is known to be true. For this representation of triples, a particular three-way factorization was developed which permits to exploit nonlocal information without explicit aggregation by collective learning. Furthermore, in [15] this approach was applied to the sizable YAGO-2 ontology, demonstrating its scalability. In [10] an additive model is described that attempts to combine the simplicity of the SUNS approach with some of the powerful features of a tensor model.

6 Conclusions

In this chapter we have argued that machine learning might become a third important way to access Web information, in addition to keyword-based search and structured querying. We have provided examples that illustrate for what kind of queries machine learning might be effective. We have discussed the machine learning approaches pursued in the work in THESEUS and LarKC, which are based on the factorization of matrices and tensors. We expect that machine learning researchers will increasingly consider the Web as a great data source for learning applications. In particular the joint exploitation of different knowledge sources like Web documents, Wikipedia, published databases as part of LOD, and other background information poses new interesting research challenges.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. The Semantic Web (2008)
2. Balduini, M., Celino, I., Dell'Aglio, D., Valle, E.D., Huang, Y., Lee, T., Kim, S.H., Tresp, V.: Reality mining on micro-post streams: Deductive and inductive reasoning for personalized and location-based recommendations. Submitted (2012)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) (2009)
4. Brickley, D., Miller, L.: The Friend of a Friend (FOAF) project. http://www.foaf-project.org/
5. Fensel, D., van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Valle, E.D., Fischer, F., Huang, Z., Kiryakov, A., il Lee, T.K., Schooler, L., Tresp, V., Wesner, S., Witbrock, M., Zhong, N.: Towards larkc: A platform for web-scale reasoning. In: ICSC (2008)
6. Huang, Y., Bundschus, M., Tresp, V., Rettinger, A., Kriegel, H.P.: Multivariate structured prediction for learning on the semantic web. In: ILP (2010)
7. Huang, Y., Nickel, M., Tresp, V., Kriegel, H.P.: A scalable kernel approach to learning in semantic graphs with applications to linked data. In: Proceedings of the 1st Workshop on Mining the Future Internet (2012)
8. Huang, Y., Tresp, V., Nickel, M., Rettinger, A., Kriegel, H.P.: A scalable approach for statistical learning in semantic graphs. Semantic Web Interoperability, Usability, Applicability (SWJ) (2012). Accepted for publication
9. Jiang, X., Huang, Y., Nickel, M., Tresp, V.: Combining information extraction, deductive reasoning and machine learning for relation prediction. In: Proceedings of the ESWC (2012)
10. Jiang, X., Tresp, V., Huang, Y., Nickel, M.: Link prediction in multi-relational graphs using additive models. In: Proceedings of International ISWC Workshop on Semantic Technologies meet Recommender Systems & Big Data (2012)
11. Jiang, X., Tresp, V., Huang, Y., Nickel, M., Kriegel, H.P.: Scalable relation prediction exploiting both intrarelational correlation and contextual information. In: Proceedings of the ECML/PKDD (2012)
12. Kann, M.G.: Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Briefing in Bioinformatics **11** (2010)
13. LarKC: The large Knowledge Collider. EU FP 7 Large-Scale Integrating Project, http://www.larkc.eu/ (2008)
14. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: Proceedings of the 28th International Conference on Machine Learning (2011)
15. Nickel, M., Tresp, V., Kriegel, H.P.: Factorizing yago: Scalable machine learning for linked data. In: Proceedings of the 21st International World Wide Web Conference (2012)
16. Tresp, V., Huang, Y., Bundschus, M., Rettinger, A.: Materializing and querying learned knowledge. In: Proceedings of the First ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (2009)
17. Tresp, V., Huang, Y., Jiang, X., Rettinger, A.: Graphical models for relations - modeling relational context. In: International Conference on Knowledge Discovery and Information Retrieval (2011)