Probabilistic Databases of Universal Schema

Limin Yao **Sebastian Riedel** **Andrew McCallum**
Department of Computer Science
University of Massachusetts, Amherst
{lmyao,riedel,mccallum}@cs.umass.edu

Abstract

In data integration we transform information from a source into a target schema. A general problem in this task is loss of fidelity and coverage: the source expresses more knowledge than can fit into the target schema, or knowledge that is hard to fit into any schema at all. This problem is taken to an extreme in information extraction (IE) where the source is natural language. To address this issue, one can either automatically learn a latent schema emergent in text (a brittle and ill-defined task), or manually extend schemas. We propose instead to store data in a probabilistic database of *universal schema*. This schema is simply the union of all source schemas, and the probabilistic database learns how to predict the cells of each source relation in this union. For example, the database could store Freebase relations and relations that correspond to natural language surface patterns. The database would learn to predict what freebase relations hold true based on what surface patterns appear, and vice versa. We describe an analogy between such databases and collaborative filtering models, and use it to implement our paradigm with probabilistic PCA, a scalable and effective collaborative filtering method.

1 Introduction

Natural language is a highly expressive representation of knowledge. Yet, for many tasks databases are more suitable, as they support more effective decision support, question answering and data mining. But given a fixed schema, any database can only capture so much of the information natural language can express, even if we restrict us to factual knowledge. For example, Freebase (Bollacker et al., 2008) captures the content of Wikipedia to some extent, but has no *criticized*(Person,Person) relation and hence cannot answer a question like "Who criticized George Bush?", even though partial answers are expressed in Wikipedia. This makes the database schema a major bottleneck in information extraction (IE). From a more general point of view, data integration always suffers from schema mismatch between knowledge source and knowledge target.

To overcome this problem, one could attempt to manually extend the schema whenever needed, but this is a time-consuming and expensive process. Alternatively, in the case of IE, we can automatically induce latent schemas from text, but this is a brittle, ill-defined and error-prone task. This paper proposes a third alternative: sidestep the issue of incomplete schemas altogether, by simply combining the relations of all knowledge sources into what we will refer to as a *universal schema*. In the case of IE this means maintaining a database with one table per natural language surface pattern. For data integration from structured sources it simply means storing the original tables as is. Crucially, the database will not only store what each source table *does* contain, it will also learn a probabilistic model about which other rows each source table *should correctly* contain.

Let us illustrate this approach in the context of IE. First we copy tables such as *profession* from a structured source (say, DBPedia). Next we create one table per surface pattern, such as *was-criticized-by* and *was-attacked-by* and fill these tables with the entity pairs that appear with this pattern in some natural language corpus (say, the NYT Corpus). At this point, our database is a simple combination of

a structured and an OpenIE (Etzioni et al., 2008) knowledge representation. However, while we insert this knowledge, we can learn a probabilistic model which is able to predict *was-criticized-by* pairs based on information from the *was-attacked-by* relation. In addition, it learns that the *profession* relation in Freebase can help disambiguate between physical attacks in sports and verbal attacks in politics. At the same time, the model learns that the natural language relation *was-criticized-by* can help predict the *profession* information in Freebase. Moreover, often users of the database will not need to study a particular schema—they can use their own expressions (say, *works-at* instead of *profession*) and still find the right answers.

In the previous scenario we could answer more questions than our structured sources alone, because we learn how to predict new Freebase rows. We could answer more questions than the text corpus and OpenIE alone, because we learn how to predict new rows in surface pattern tables. We could also answer more questions than in Distant Supervision (Mintz et al., 2009), because our schema is not limited to the relations in the structured source. We could even go further and import additional structured sources, such as Yago (Hoffart et al., 2012). In this case the probabilistic database would have integrated, and implicitly aligned, several different data sources, in the sense that each helps predict the rows of the other.

In this paper we present results of our first technical approach to probabilistic databases with universal schema: collaborative filtering, which has been successful in modeling movie recommendations. Here each entity tuple explicitly "rates" source tables as "I appear in it" or "I don't", and the recommender system model predicts how the tuple would "rate" other tables—this amounts to the probability of membership in the corresponding table. Collaborative filtering provides us with a wide range of scalable and effective machine learning techniques. In particular, we are free to choose models that use no latent representations at all (such as a graphical model with one random variable per database cell), or models with latent representations that do not directly correspond to interpretable semantic concepts. In this paper we explore the latter and use a probabilistic generalization to PCA for recommendation.

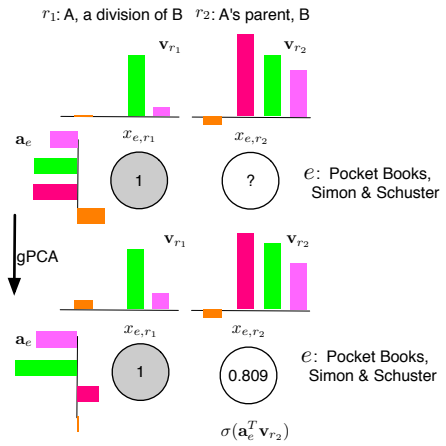

Figure 1: gPCA re-estimates the representations of two relations and a tuple with the arrival of an observation $r_1(e)$. This enables the estimation of the probability for unseen fact $r_2(e)$. Notice that both tuple and relation components are re-estimated and can change with the arrival of new observations.

In our experiments we integrate Freebase data and information from the New York Times Corpus (Sandhaus, 2008). We show that our probabilistic database can answer questions neither of the sources can answer, and that it uses information from one source to improve predictions for the other.

2 Generalized PCA

In this work we concentrate on a set of binary source relations $\mathcal{R}$, consisting of surface patterns as well as imported tables from other databases, and a set of entity pairs $\mathcal{E}$. We introduce a matrix $\mathbf{X}$ where each cell $x_{e,r}$ is a binary random variable indicating whether $r(e)$ is true or not. The upper half of figure 1 shows two cells of this matrix, based on the relations r_1 ("*a-division-of*") and r_2 ("'*s-parent,*") and the tuple e (Pocket Books, Simon&Schuster). Generally some of the cells will be observed (such as x_{e,r_1}) while others will be not (such as x_{e,r_2}).

We employ a probabilistic generalization of Principle Component Analysis (gPCA) to estimate the probabilities $P(r(e))$ for every non-observed fact $r(e)$ (Collins et al., 2001). In gPCA we learn a k-dimensional feature vector representation $\mathbf{v}_r$ for each relation (column) r, and a k-dimensional feature vector representation $\mathbf{a}_e$ for each entity pair e.

Figure 1 shows example vectors for both rows and columns. Notice that these vectors do not have to be positive nor sum up to one. Given these representations, the probability of $r(e)$ being true is given by the logistic function $\sigma(\theta) = \frac{1}{1+\exp(-\theta)}$ applied to the dot product $\theta_{r,e} \triangleq \mathbf{a}_e^\intercal \mathbf{v}_r$. In other words, we represent the matrix of parameters $\mathbf{\Theta} \triangleq (\theta_{r,e})$ using a low-rank approximation $\mathbf{AV}$ where $\mathbf{A} = (\mathbf{a}_e)_{e \in \mathcal{E}}$ and $\mathbf{V} = (\mathbf{v}_r)_{r \in \mathcal{R}}$.

Given a set of observed cells, gPCA estimates the tuple feature representations $\mathbf{A}$ and the relation feature representations $\mathbf{V}$ by maximizing the log-likelihood of the observed data. This can be done both in batch mode or in a more incremental fashion. In the latter we observe new facts (such as $r_1(e)$ in Figure 1) and then re-estimate $\mathbf{A}$ and $\mathbf{V}$. In Figure 1 we show what this means in practice. In the upper half we see the currently estimated representation $\mathbf{v}_{r_1}$ and $\mathbf{v}_{r_2}$ of r_1 and r_2, and a random initialization for the representation $\mathbf{a}_e$ of e. In the lower half we take the observation $r_1(e)$ into account and re-estimate $\mathbf{a}_e$ and $\mathbf{v}_{r_1}$. The new estimates can then be used to calculate the probability $\sigma(\mathbf{a}_e^\intercal \mathbf{v}_{r_2})$ of $r_2(e)$.

Notice that by incorporating new evidence for a given row, both entity and relation representations can improve, and hence beliefs across the whole matrix. In this sense, gPCA performs a form of joint or global inference. Likewise, when we observe several active relations for a new tuple, the model will increase the probabilistic association between theses relations and, transitively, also previously associated relations. This gives gPCA a never-ending-learning quality. Also note that it is easy to incorporate entity representations into the approach, and model selectional preferences. Likewise, we can easily add posterior constraints we know to hold across relations, and learn from unlabeled data.

3 Related Work

We briefly review related work in this section. Open IE (Etzioni et al., 2008) extracts how entities and their relations are actually mentioned in text, but does not predict how entities could be mentioned otherwise and hence suffer from reduced recall. There are approaches that learn synonym relations between surface patterns (Yates and Etzioni, 2009; Pantel et al., 2007; Lin and Pantel, 2001; Yao et al., 2011) to overcome this problem. Fundamentally, these methods rely on a *symmetric* notion of synonymy in which certain patterns are assumed to have the same meaning. Our approach rejects this assumption in favor of a model which learns that certain patterns, or combinations thereof, *entail* others in one direction, but not necessarily the other.

Methods that learn rules between textual patterns in OpenIE aim at a similar goal as our proposed gPCA algorithm (Schoenmackers et al., 2008; Schoenmackers et al., 2010). Such methods learn the structure of a Markov Network, and are ultimately bounded by limits on tree-width and density. In contrast, the gPCA learns a latent, although not necessarily interpretable, structure. This latent structure can express models of very high tree-width, and hence very complex rules, without loss in efficiency. Moreover, most rule learners work in batch mode while our method continues to learn new associations with the arrival of new data.

4 Experiments

Our work aims to predict new rows of source tables, where tables correspond to either surface patterns in natural language sources, or tables in structured sources. In this paper we concentrate on binary relations, but note that in future work we will use unary, and generally n-ary, tables as well.

4.1 Unstructured Data

The first set of relations to integrate into our universal schema comes from the surface patterns of 20 years of New York Times articles (Sandhaus, 2008). We preprocess the data similarly to Riedel et al. (2010). This yields a collection of entity mention pairs that appear in the same sentence, together with the syntactic path between the two mentions.

For each entity pair in a sentence we extract the following surface patterns: the dependency path which connects the two named entities, the words between the two named entities, and the context words of the two named entities. Then we add the entity pair to the set of relations to which the surface patterns correspond. This results in approximately 350,000 entity pairs in 23,000 relations.

Table 1: GPCA fills in new predicates for records

Relation	<-subj<-**own**->obj->**perc**.>prep->**of**->obj->	<-subj<-**criticize**->obj->
Obs.	Time Inc., American Tel. and Comms.	Bill Clinton, Bush Administration
New	United States, Manhattan	Mr. Forbes, Mr. Bush
	Campeau, Federated Department Stores	Mr. Dinkins, Mr. Giuliani
	Volvo, Scania A.B.	Mr. Badillo, Mr. Bloomberg

4.2 Structured Data

The second set of source relations stems from Freebase. We choose those relations that hold for entity pairs appearing in the NYT corpus. This adds 116 relations to our universal schema. For each of the relations we import only those rows which correspond to entity tuples also found in the NYT corpus. In order to link entity mentions in the text to entities in Freebase, we follow a simple string-match heuristic.

4.3 Experimental Setup and Training

In our experiments, we hold out some of the observed source rows and try to predict these based on other observed rows. In particular, for each entity pair, we traverse over all source relations. For each relation we throw an unbiased coin to determine whether it is observed for the given pair. Then we train a gPCA model of 50 components on the observed rows, and use it to predict the unobserved ones. Here a pair e is set to be in a given relation r if $P(r(e)) > 0.5$ according to our model. Since we generally do not have observed negative information,[1] we sub-sample a set of negative rows for each relation r to create a more balanced training set.

We evaluate recall of our method by measuring how many of the true held out rows we predict. We could use a similar approach to measure precision by considering each positive prediction to be a false positive if the observed held-out data does not contain the corresponding fact. However, this approach underestimates precision since our sources are generally incomplete. To overcome this issue, we use human annotations for the precision measure. In particular, we randomly sample a subset of entity pairs and ask human annotators to assess the predicted positive relations of each.

[1] Just because a particular e has not yet been seen in particular relation r we cannot infer that $r(e)$ is false.

4.4 Integrating the NYT Corpus

We investigate how gPCA can help us answer questions based on only single data source: the NYT Corpus. Table 1 presents, for two source relations (aka surface patterns), a set of observed entity pairs (Obs.) and the most likely inferred entity pairs (New). The table shows that we can answer a question like "Who owns percentages of Scania AB?" even though the corpus does not explicitly contain the answer. In our case, it only contains *"buy-stake-in*(VOLVO,SCANIA AB)."

gPCA achieves about 49% recall, at about 67% precision. Interestingly, the model learns more than just paraphrasing. Instead, it captures some notion of entailment. This can be observed in its asymmetric beliefs. For example, the model learned to predict "*professor-at*(K.BOYLE, OHIO STATE)" based on "*historian-at*(KEVIN BOYLE, OHIO STATE)" but would not make the inference "*historian-at*(R.FREEMAN,HARVARD)" based on "*professor-at*(R.FREEMAN,HARVARD)."

4.5 Integrating Freebase

What happens if we integrate additional structured sources into our probabilistic database? We observe that by incorporating Freebase tables in addition to the NYT data we can improve recall from 49% to 52% on surface patterns. The precision also increases by 2%.

Table 2 sums the results and also gives an example of how Freebase helps improve both precision and recall. Without Freebase, the gPCA predicts that Maher Arar was arrested in Syria—primarily because he lived in Syria and the NYT often talks about arrests of people in the city they live in[2]. After learning *placeOfBirth*(ARAR,SYRIA) from Freebase, the gPCA model infers *wasBornIn*(ARAR,SYRIA) as well as *grewUpIn*(ARAR,SYRIA).

[2] In fact, he was arrested in US

Table 2: Relation predictions w/o Freebase.

	without Freebase	with Freebase
Prec.	0.687	0.666
Rec.	0.491	0.520
E.g.	M. Arar, Syria (Freebase: placeOfBirth)	
Pred.	A was arrested in B	A was born in B
	A appeal to B	A grow up in B
	A, who represent B	A's home in B

5 Conclusion

In our approach we do not design or infer new relations to accommodate information from different sources. Instead we simply combine source relations into a universal schema, and learn a probabilistic model to predict what other rows the sources could contain. This simple paradigm allows us to perform data alignment, information extraction, and other forms of data integration, while minimizing both loss of information and the need for schema maintenance.

At the heart of our approach is the hypothesis that we should concentrate on building models to predict source data—a relatively well defined task—as opposed to models of semantic equivalence that match our intuition. Our future work will therefore investigate such predictive models in more detail, and ask how to (a) incorporate relations of different arities, (b) employ background knowledge, (c) optimize the choice of negative data and (d) scale up both in terms of rows and tables.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and the University of Massachusetts and in part by UPenn NSF medium IIS-0803847. We gratefully acknowledge the support of Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA, AFRL, or the US government.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, New York, NY, USA. ACM.

Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. 2001. A generalization of principal component analysis to the exponential family. In *Proceedings of NIPS*.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2012. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*.

Dekang Lin and Patrick Pantel. 2001. DIRT - discovery of inference rules from text. In *Knowledge Discovery and Data Mining*, pages 323–328.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL '09)*, pages 1003–1011. Association for Computational Linguistics.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning Inferential Selectional Preferences. In *Proceedings of NAACL HLT*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10)*.

Evan Sandhaus, 2008. *The New York Times Annotated Corpus*. Linguistic Data Consortium, Philadelphia.

Stefan Schoenmackers, Oren Etzioni, and Daniel S. Weld. 2008. Scaling textual inference to the web. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–88, Morristown, NJ, USA. Association for Computational Linguistics.

Stefan Schoenmackers, Oren Etzioni, Daniel S. Weld, and Jesse Davis. 2010. Learning first-order horn clauses from web text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1088–1098,

Stroudsburg, PA, USA. Association for Computational Linguistics.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical methods in natural language processing (EMNLP '11)*, July.

Alexander Yates and Oren Etzioni. 2009. Unsupervised methods for determining object and relation synonyms on the web. *Journal of Artificial Intelligence Research*, 34:255–296.