

Research Statement - Ronen Feldman

The information age has made it easy to store large amounts of data. The proliferation of documents available on the Web, on corporate intranets, on news wires, and elsewhere is overwhelming. However, while the amount of data available to us is constantly increasing, our ability to absorb and process this information remains constant. Search engines only exacerbate the problem by making more and more documents available in a matter of a few key strokes. Text Mining is a new and exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, NLP, IR and knowledge management. Text Mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (distribution analysis, clustering, trend analysis, association rules etc) and visualization of the results. My research evolves around the various components of text mining. In the following sections I will describe the various research activities that I have done in the recent years and plans for future research. My main motto in research is the combination of theory and practice and indeed in each of the following areas we have developed a complete theory and proved that it actually works in practice by implementing a large scale system based on the theory.

Hybrid Information Extraction

The knowledge engineering (mostly rule-based) systems traditionally were the top performers in most IE benchmarks, such as MUC (Chinchor, Hirschman et al. 1994), ACE and the KDD CUP (Yeh and Hirschman 2002). Recently, though, the machine learning systems became state of the art, especially for simpler tagging problems, such as named entity recognition (Bikel, Miller et al. 1997) or field extraction (McCallum, Freitag et al. 2000). Still, the knowledge-engineering approach retains some of its advantages. It is focused around manually writing patterns to extract the entities and relations. The patterns are naturally accessible to human understanding and can be improved in a controllable way. Whereas improving the results of a pure machine-learning system would require providing it with additional training data. However, the impact of adding more data soon becomes infinitesimal while the cost of manually annotating the data grows linearly. We have developed a hybrid entities-and relations-extraction system, which combines the power of knowledge-based and statistical machine-learning approaches. The system is based on stochastic context-free grammars. It is called TEG, for trainable extraction grammar. The rules for the extraction grammar are written manually, while the probabilities are trained from an annotated corpus. The powerful disambiguation ability of PCFGs allows the knowledge engineer to write very simple and naive rules while retaining their power, thus greatly reducing the required labor. In addition, the size of the needed training data is considerably smaller than the size of the training data needed for pure machine-learning systems (for achieving comparable accuracy results). Furthermore, the tasks of rule writing and corpus annotation can be balanced against each other.

Plans for Future Research. TEG is based on the combination of HMM and CFG. We conjecture that by combining a stronger machine learning algorithm such as CRF, RMM or MIRA with CFG we can create a more powerful hybrid solution. It is our goal to create a hybrid solution that will solve most of the problems that we encountered while trying to develop IE modules with TEG. The most severe of those

was the lack of a clear methodology for building accurate IE modules using the CFG formalism of TEG.

Unsupervised Web Extraction

Information Extraction (IE) (Riloff 1993; Cowie and Lehnert 1996; Grishman 1996; Grishman 1997; Kushmerick, Weld et al. 1997; Freitag 1998; Freitag and McCallum 1999; Soderland 1999) is the task of extracting factual assertions from text.

Most IE systems rely on knowledge engineering or on machine learning to generate *extraction patterns* – the mechanism that extracts entities and relation instances from text. In the machine learning approach, a domain expert labels instances of the target relations in a set of documents. The system then learns extraction patterns, which can be applied to new documents automatically.

Both approaches require substantial human effort, particularly when applied to the broad range of documents, entities, and relations on the Web. In order to minimize the manual effort necessary to build Web IE systems, we have designed and implemented URES (Unsupervised Relation Extraction System). URES takes as input the names of the target relations and the types of their arguments. It then uses a large set of unlabeled documents downloaded from the Web in order to learn the extraction patterns.

URES is most closely related to the KnowItAll system developed at University of Washington by Oren Etzioni and colleagues (Etzioni, Cafarella et al. 2005), since both are unsupervised and both leverage relation-independent extraction patterns to automatically generate seeds, which are then fed into a pattern-learning component. KnowItAll is based on the observation that the Web corpus is highly redundant. Thus, its selective, high-precision extraction patterns readily ignore most sentences, and focus on sentences that indicate the presence of relation instances with very high probability.

In contrast, URES is based on the observation that, for many relations, the Web corpus has *limited redundancy*, particularly when one is concerned with less prominent instances of these relations (e.g., the acquisition of Austria Tabak). Thus, URES utilizes a more expressive extraction pattern language, which enables it to extract information from a broader set of sentences. URES relies on a sophisticated mechanism to assess its confidence in each extraction, enabling it to sort extracted instances, thereby improving its recall without sacrificing precision.

Our main contributions are as follows:

1. We introduced the first domain-independent system to extract relation instances from the Web with both high precision and high recall.
2. We showed how to minimize the human effort necessary to deploy URES for an arbitrary set of relations, including automatically generating and labeling positive and negative examples of the relation.
3. We performed an experimental comparison between URES and the state-of-the-art KnowItAll system, and showed that URES can double or even triple the recall achieved by KnowItAll for relatively rare relation instances.

Plans for Future Research. This research area is very promising and we expect that it can revolutionaries the whole area of text mining. We have many research goals that we want to achieve within the next 5-10 years.

1. Integrate anaphora resolution into URES so that we can extract relations that are spread across multiple sentences.
2. Integrate a NER component into URES (rule based, CRF, RMM) to test how precision can be improved vs. simple NP extraction.
3. Move to a more powerful pattern language that will include more linguistic features.
4. Move from binary predicates to n-ary predicates (such as management change, earning announcements, etc.)
5. Utilize clustering techniques to learn families of relations simultaneously (like family relations, business relations between people)
6. Find techniques to boost the recall of unsupervised web extraction while still maintaining the high precision.
7. Utilize the web to validate the results of URES
8. Utilize URES like techniques for classic information extraction
9. Use pattern matching algorithms to reduce the computational complexity of URES.

Visual Information Extraction

Most information extraction systems simplify the structure of the documents they process by ignoring much of the visual characteristics of the document, e.g. font type, size and location, and process the text as a linear sequence. This allows the algorithms to focus on the semantic aspects of the document. However, valuable information is lost. Consider, for example, an article in a scientific journal. The title is readily recognized based on its special font and location, but less so based on its semantic content, which may be similar to the section headings. Similarly, for the author names, section headings, running title, etc. Thus, much important information is provided by the visual layout of the document. We have developed an information extraction system that is based solely on the visual characteristics of the document, and have shown that this visual information alone is sufficient to provide high accuracy extraction, for specific fields (e.g. the title, author names, publication date, etc.).

We developed a general algorithm which allows to perform the IE task based on the visual layout of the document. The algorithm employs a machine learning approach whereby the system is first provided with a set of training documents in which the desired fields are manually tagged. Based on these training examples the system automatically learns how to find the corresponding fields in future documents.

Problem Formulation. A document D is a set of *primitive elements* $D = \{e_1, \dots, e_n\}$. A primitive element can be a character, a line, or any other visual object, depending on the document format. A primitive element may have any number of visual attributes, such as font size and type, physical location, etc. The *bounding box* attribute, which provides the size and location of the bounding box of the element, is assumed to be available for all primitive elements. We define an *object* in the document to be any set of primitive elements. The *Visual Information Extraction (VIE) task* is as follows. We are provided with a set of *target fields* $F = \{f_1, \dots, f_k\}$, to be extracted, and a set of *training documents* $T = \{T_1, \dots, T_m\}$ wherein all occurrences of the target fields are

tagged. Specifically, for each target field f and training document T , we are provided with the object $f(T)$ of T that is of type f ($f(T) = ;$ if f does not appear in T). The goal is that when presented with an un-tagged query document Q , to correctly tag the occurrences of the target fields that exist in Q (not all target fields need be present in each document).

Results. We have developed a general framework and algorithm for the VIE task. We have shown that the VIE task can be decomposed into two subtasks. First, for each document (both training and query) we must group the primitive elements into meaningful objects (e.g. lines, paragraphs, etc.), and establish the hierarchical structure among these objects. Then, in the second stage, the structure of the query document is compared with those of the training documents to find the objects corresponding to the target fields. We have also shown how to improve the results by introducing the notion of *templates* which are groups of training documents with a similar layout (say, articles from the same journal). Using templates we can identify the essential features of the page layout, ignoring particularities of any specific document. We implemented the system for a VIE task on a set of documents containing financial analyst reports. The documents were in PDF format. Target fields included the title, authors, publication dates, and others.

Plans for Future Research. Clearly, our visual approach also has its limitations. First and foremost, the visual approach can only capture fields with distinct visual characteristics, such as the title, authors, publication date, etc. Semantic elements mentioned within the running text, such as people names, locations, etc., clearly cannot be detected by the visual approach. In addition, the learning process that we used only works for features and structures that have a relatively high level of consistency among documents, such as title, author, etc. The method would be less applicable to structures with a high level of variations between documents. Ultimately, we believe that a complete solution for information extraction should make use of the entire spectrum of available information: semantic, syntactic and visual. In such a system, our visual approach would be one of the components in a combined, integrated approach. It is one of our main goals within the next 5 years to develop such a system.

Additional Areas of Active Research

- A visual query language (and efficient query execution engine) for link analysis
- A visual text mining environment
- An integrated system for email mining (using the Enron email repository as a test case)
- Integration of Information Extraction into the Semantic Web framework (OWL, RDF, etc.)
- Information extraction in Hebrew and Arabic (developed for the Israeli MOD)
- Automatic identification and correction of annotation errors (for machine learning based IE)
- Automatic construction of Knowledge Bases from the PubMed database (in collaboration with UPenn's CS Dept and Medical School)
- An IE based search engine (in collaboration with U of Washington, Seattle)
- Temporal Text Mining Environment
- Text Mining of Chat Rooms and Messenger Logs.

References

- Bikel, Daniel M., Miller, Scott, Schwartz, Richard and Weischedel, Ralph (1997). Nymble: a high-performance learning name-finder. Proceedings of ANLP-97: 194-201.
- Chinchor, Nancy, Hirschman, Lynnette and Lewis, David (1994). "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)." Computational Linguistics **3**(19): 409-449.
- Cowie, J. and Lehnert, W. (1996). "Information Extraction." Communications of the Association of Computing Machinery **39**(1): 80-91.
- Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D. and Yates, A. (2005). "Unsupervised named-entity extraction from the Web: An experimental study." Artificial Intelligence.
- Freitag, Dayne (1998). Machine Learning for Information Extraction in Informal Domains. Computer Science Department. Pittsburgh, PA, Carnegie Mellon University: 188.
- Freitag, Dayne and McCallum, Andrew Kachites (1999). Information extraction with HMMs and shrinkage. Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction.
- Grishman, R. (1996). The role of syntax in Information Extraction. Advances in Text Processing: Tipster Program Phase II, Morgan Kaufmann.
- Grishman, Ralph (1997). Information Extraction: Techniques and Challenges. SCIE: 10-27.
- Kushmerick, Nickolas, Weld, Daniel S. and Doorenbos, Robert B. (1997). Wrapper Induction for Information Extraction. Intl. Joint Conference on Artificial Intelligence (IJCAD): 729-737.
- McCallum, Andrew, Freitag, Dayne and Pereira, Fernando (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA: 591-598.
- Riloff, Ellen (1993). Automatically Constructing a Dictionary for Information Extraction Tasks. In Proceedings of the Eleventh National Congress on Artificial Intelligence, AAAI Press / MIT Press.
- Soderland, Stephen (1999). "Learning Information Extraction Rules for Semi-Structured and Free Text." Machine Learning **34**(1-3): 233-272.
- Yeh, A and Hirschman, L (2002). "Background and Overview for KDD Cup 2002 Task 1: Information Extraction from Biomedical Articles." KDD Explorations **4**(2): 87-89.