Exploring the Corporate Ecosystem with a Semi-Supervised Entity Graph

Hassan H. Malik
Thomson Reuters
New York, NY, USA
hassan.malik
@thomsonreuters.com

Ian MacGillivray
Thomson Reuters
New York, NY, USA
ian.macgillivray
@thomsonreuters.com

Måns Olof-Ors
Thomson Reuters
Baar, Switzerland
mans.olof-ors
@thomsonreuters.com

Siming Sun
Thomson Reuters
New York, NY, USA
siming.sun
@thomsonreuters.com

Shailesh Saroha
Thomson Reuters
New York, NY, USA
shailesh.saroha
@thomsonreuters.com

ABSTRACT

Investment decisions in the financial markets require careful analysis of information available from multiple data sources. In this paper, we present Atlas, a novel entity-based information analysis and content aggregation platform that uses heterogeneous data sources to construct and maintain the "ecosystem" around tangible and logical entities such as organizations, products, industries, geographies, commodities and macroeconomic indicators. Entities are represented as vertices in a directed graph, and edges are generated using entity co-occurrences in unstructured documents and supervised information from structured data sources. Significance scores for the edges are computed using a method that combines supervised, unsupervised and temporal factors into a single score.

Important entity attributes from the structured content and the entity neighborhood in the graph are automatically summarized as the entity "fingerprint". A highly interactive user interface provides exploratory access to the graph and supports common business use cases.

We present results of experiments performed on five years of news and broker research data, and show that Atlas is able to accurately identify important and interesting connections in real-world entities. We also demonstrate that Atlas entity fingerprints are particularly useful in entity similarity queries, with a quality that rivals existing human-maintained databases.

Categories and Subject Descriptors

E.1 [**DATA STRUCTURES**]: Graphs and networks; H.4.2 [**Information Systems Applications**]: Decision support

General Terms

Algorithms, Design, Experimentation, Performance

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

1. INTRODUCTION AND MOTIVATION

Financial analysts and investors often rely on news stories, press releases and legal documents to make critical investment decisions. However, facts or events reported in a single document are rarely sufficient for an investor to make an informed decision. Therefore, investors must conduct additional research that typically involves browsing and searching multiple structured and unstructured databases. Since the windows of opportunity in financial markets are typically short, investors often face a difficult tradeoff between conducting thorough research and making timely decisions.

We observe that an interesting property of new documents is that they sometimes substantially increase the significance of existing documents. This is because significant new facts or events that are related to an entity tend to revive interest in older information about that entity. Often, this 'boost' in information significance propagates to entities that may not even be mentioned in the new document. Due to this phenomenon, seemingly unrelated news stories may have substantial financial implications on real-life assets.

For example, the initial news alerts reporting the recent political unrest in Egypt merely reported the events. Since Egypt is one of the largest importers of wheat in the world, the unrest soon negatively impacted wheat prices in the international markets, which in turn may have impacted the financial standing of some of the commodity-based investment vehicles and private and publicly traded companies with significant exposure to wheat, in addition to companies with exposure to the country or the region in general. Again, existing databases (and news archives) are likely to contain all information needed to analyze possible implications of the unrest, but an average investor may not even be able to find all relevant information before the markets start responding to the events.

In this paper, we present Atlas, a novel entity-based information analysis and content aggregation platform that uses heterogeneous structured and unstructured data sources to construct and maintain the 'ecosystem' around tangible and logical entities such as organizations, products, industries, geographies, commodities and macroeconomic indicators. Atlas uses a directed graph as the main data structure where

entities are represented as vertices, and edges are generated using entity co-occurrences in unstructured documents and supervised information from structured data sources. To facilitate efficient access to underlying content and to allow for temporal analysis, vertices and edges store pointers to relevant unstructured and structured content and metadata.

Since each entity may be connected to a large number of other entities, Atlas computes significance scores for edges using a novel method that combines supervised, unsupervised and temporal factors into a single score. In addition, Atlas automatically summarizes important entity attributes from the structured content and the entity neighborhoods in the graph into entity 'fingerprints'. As we show in Section 6.2.1, these fingerprints are particularly useful in entity similarity queries. Finally, a highly interactive user interface provides exploratory access to the graph and supports common business use cases.

The rest of this paper is organized as follows. We begin with a discussion of related work (Section 2) and provide an overview of Atlas system architecture (Section 3). We then describe the graph construction in detail (Section 4) providing information on the data sources used (Section 4.1), issues and challenges encountered (Section 4.2.1), the computation of significance scores for edges (Section 4.3), the generation of entity fingerprints (Section 4.4), and the characteristics of the generated graph (Section 4.5). Next, we provide an overview of the user interfaces (Section 5) and empirically evaluate the graph's performance (Section 6). We finally conclude and discuss ideas for future work (Section 7).

2. RELATED WORK

Our research relates to existing work in entity extraction and disambiguation, entity-relation graphs, news event tracking and news document chaining, standardized analysis and management of unstructured information, and cross-database visualization, along with other research areas in information retrieval, database management, knowledge management and machine learning. We discuss a few representative methods and systems here.

Entity extraction and disambiguation serves as the basic building block in our system. Earlier entity extraction systems were mostly rule-based [11, 24, 15] whereas statistical methods [28, 12, 10] gained more popularity in recent years. These methods convert the entity extraction task to a problem of decomposing the unstructured text, and then labeling various parts of the decomposition [25]. Common methods of decomposing unstructured text include splitting the unstructured text along a predefined set of delimiters or into word chunks using NLP-based methods. Statistical methods such as Hidden Markov Models [1, 3], Maximum-Entropy-based methods [4], and Conditional Random Fields [18] are popular for labeling the decomposed text. Since multiple extracted entities may represent the same physical entity (e.g., International Business Machines and IBM both represent the same company,) extracted entities must be disambiguated to avoid unnecessary duplication.

Bunescu and Pasca [6] used Wikipedia to train a disambiguation SVM kernel whereas Cucerzan [9] proposed a disambiguation process that focuses on maximizing the agreement between the document context, Wikipedia context, and the category tags associated with the candidate entities. In contrast, Hassell et al. [14] used an Ontology as background knowledge for entity disambiguation.

Entity-Relation (ER) graphs have been proposed to model entity relationships. Chakrabarti et al. [8] used ER graphs to represent personal information networks, where nodes represented entities such as organizations, people, places, events, projects etc and edges represented explicit or probabilistic relationships obtained by parsing unstructured text. A proximity-based query language supported queries on the graph. In follow-up work [7], they used a PageRank-like method to improve the query execution performance. Minkov and Cohen [22] also used ER graphs to model personal information, and used finite graph walks to induce a measure of entity similarity and to facilitate searching the graph.

News event tracking and finding connections between news stories is another active research area that is related to our work. Nallapati et al. [23] studied the problem of recognizing events and their dependencies in news stories. They generated a graph structure by discovering sub-clusters in news events and organizing them by their dependencies. In a similar approach, Mei and ChengXiang [21] discovered latent themes from text, constructed an evolution graph of themes and used HMMs to analyze the life cycle of these themes. Recently, Shahaf and Guestrin [27] studied the problem of finding coherent chains that connect a pair of news stories. They formalized the notion of story coherence as a linear program, and used a bipartite graph to measure the influence of a document on other documents. They also proposed methods to find and evaluate coherent chains.

Managing, analyzing and visualizing data from many structured and unstructured data sources is often challenging. Ferrucci and Lally [13] proposed a middleware that provides standardized interfaces to acquire, analyze and access unstructured data. Their framework supports document-level and collection-level analysis and enables semantic searches and standardized access to resulting metadata. Lieberman et al. [19] proposed a query-based approach to visualize and explore heterogenous biomedical databases. They modeled database records as nodes in an ER graph, and used edges to link related records. Keyword-based queries return an initial set of nodes, and the user is then able to explore the links to records from multiple databases in a unified way.

Existing entity centric content aggregation systems are either community maintained or use proprietary methods. Freebase[1] is a community maintained entity graph that contains information on about 20 million[2] unique entities. Entities are associated with one or more types, and may have additional properties: they are stored as nodes in a graph database and links represent relationships between entities. DBpedia [2] is another community effort that aims to extract structured information about entities from Wikipedia and make it accessible on the web. Various facts about entities, relationships to other sources, classifications in multiple concept hierarchies and data-level links to other web data sources are also maintained. The OKKAM [5] project aims to create a web-scalable entity name system to enable entity-centric information integration, and Quid[3] uses proprietary methods to map the world's technologies with a goal of helping businesses identify their next strategic opportunities.

Similarly to the community maintained systems, Atlas utilizes supervised information about entities and their relationships, where available. However, our work differs from

[1]http://www.freebase.com
[2]At the time of writing
[3]http://www.quid.com

these systems in that we do not wholly rely on supervised information to add entities to the graph, or to establish connections. Instead, we use any available entity extraction and resolution system to find entities in unstructured documents, establish entity connections based on co-occurrences in these documents, and apply a novel method to compute the significance of these connections. In addition, we may then use any available supervised information, entity-to-document mappings, and the entity's neighborhood in the graph to automatically summarize important attributes as an entity 'fingerprint,' resulting in a significantly more scalable system.

3. SYSTEM ARCHITECTURE

The underlying data structure of the Atlas system is a complex directed graph, where each vertex represents an entity and each edge a connection between entities. The overall Atlas system, however, also includes the creation of this graph (Section 4) and a set of query interfaces that support a number of real-world use cases (Section 5). Here we introduce the terminology used in the rest of this paper and present the architecture of our system, along with tools and technologies used in various system components, and describe in more detail the fundamentals of the graph itself.

3.1 The Entity-Centric Model

The key component in the Atlas system is that of an entity. An entity in this model may be a tangible reality, such as a person or a commodity; or an intangible concept, such as inflation, or a war. Both vertices and edges have properties and references to relevant documents, and a single edge may represent multiple types of relationships between a source and a target vertex. The entity types used in the version of Atlas presented here are organizations, geographies, products, industries, commodities and macroeconomic indicators. Further entity types available in alternate versions of Atlas include people, technologies, facilities and media.

Entities may either be *validated* or *discovered*. *Validated* entities are mapped to a known entity from a human-maintained structured data source whereas *discovered* entities are solely obtained from unstructured data sources.

3.2 The Document-Powered Approach

The fundamental principle behind Atlas is that if two entities co-occur in a document, then a relationship between them is present. Two key caveats to this assumption must be fully understood in order to faithfully implement it: firstly, it must be understood that the simple appearance of a text string matching an entity in a document does not mean that the document is *about* that entity; secondly, we must realize that whilst, on a per-document basis, this is likely to result in false-positive relationships, that over the aggregate of millions of documents, these false positives will have a negligible impact on the usefulness of the Atlas graph, as they are likely to receive low significance scores (Section 4.3).

Thus, Atlas takes in a wide range of unstructured documents (Section 4.1) and the graph is built using the occurrences of entities in these documents as raw data, combined with any structured information that may also be available. The core of the resulting system therefore loosely relates to the semi-supervised learning paradigm.

3.3 Accessing the graph

The majority of queries to the Atlas graph specify one or more entities as a parameter along with additional information to fine tune the returned results. The queries are answered by analyzing (a) the query entities' intrinsic properties (such as a person's birth-date or an organization's country of incorporation) (b) the query entities' *fingerprint*, an abstraction of the entity's properties (Section 4.4) and (c) up to degree-2 neighbors of the query entities in the graph.

Atlas makes use of as many types of evidence as are available, but is able to provide meaningful answers when nothing more than the name of an entity and a set of documents in which the name appears are available. The more complete the data available, the more precise the answers returned are, but much may be inferred about an entity from its neighbors in the graph.

3.4 Implementation

The main programming language used in the development of the Atlas system is Java. The graph is loaded in memory as a runtime Java object, and persisted to the file system as a serialized Java object. All information needed to construct the graph from scratch is stored in a relational database. Due to the scale and complexity of the application, a number of open-source libraries were used in developing the system. Amongst there were: Apache Lucene and XML Beans, Jetty and VTD-XML.

The Atlas system is deployed as a set of services, accessed by Flash or HTML-based client applications (Section 5), or programmatically by a number of products and internal tools that master and update content for our organization. Requests and responses are sent using either standard HTTP, XML over HTTP, or batched using Google's Protocol Buffers[4]. Data partitioning is used to allow vertical and horizontal scalability, i.e., the graph is distributed between multiple nodes in a de-centralized peer network, each of which owns a subset of documents or entities. Protocol Buffers are used for inter-node communication to service user requests in the distributed graph (Figure 1).

4. GRAPH CONSTRUCTION

Atlas graph construction (Figure 2) begins by selecting suitable documents from all data sources (Section 4.1) and

[4]http://code.google.com/p/protobuf

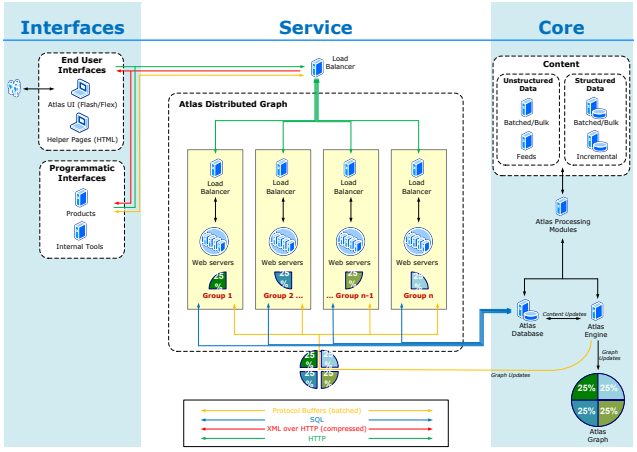


Figure 1: Distributed Implementation

preprocessing these documents (Section 4.1.1). The selected documents are then processed to extract entities and to resolve these against structured data sources (Section 4.2).

At this stage, links – which are in essence, the edges of the graph – between these resolved entities are created, and the significance of these links is established (Section 4.3). As a final step, an intrinsic representation of each entity that we call its *fingerprint* is then created (Section 4.4) using available supervised information, entity to document mappings, and the entity neighborhood in the newly established graph.

4.1 Data Sources and Preprocessing

3.8 million news stories from the Reuters News Service published between December 2005 and October 2010, and 110,000 broker research documents from a number of leading financial institutions from the same time period were used to build the graph evaluated in this paper. Multiple human-maintained, structured databases available within our organization or from on web were also used. The "Product Master" database was constructed from Freebase, and contains about 8,000 product names. The TRCS[5] database contained structured information relating to the news documents. For organizations, "Organization Authority" contains corporate, industrial and geographical information for public and private companies; "Business Sectors" maps companies to predefined industries; and "Competitors," contains a list of competitors for a small fraction of public companies.

4.1.1 Pre-processing

News is written in takes (incremental updates), each of which is published as soon as it is ready. To avoid duplication, and to allow the users to correlate all takes of a story, journalists use the same unique identifier for all takes of a story. However, there are cases where the same story is filed under multiple identifiers, causing duplications. Since these duplications may inflate the statistics, we use simple heuristics such as headline matches and cosine similarity within a time window to drop these duplicates.

In a separate problem, news and broker research documents are heavily sprinkled with boilerplate data — standard terms and phrases which often mention company names and can cause misleading connections to be formed. In our removal algorithm, we hash each period and/or newline-delimited sentence that occurs in the entire dataset, and eliminate those which occur more than a given number of standard deviations from the mean. One of the most common pieces of Boilerplate seen by Atlas is provided below:

"This information is provided by RNS The company news service from the London Stock Exchange"

After these steps, any documents which are mostly empty were removed. Further, any documents which mention more than a given number of entities n were also not considered when building our graphs, as for high values of n no examples were found where n! / 2! (n-2)! meaningful connections were actually present in the document. The value of n used in our experiments was empirically selected as 20.

[5]TRCS (Thomson Reuters Classification System) is a manually applied journalistic classification scheme, which was available to us for news documents, with tags such as OILG for a news story about Oil and Gas.

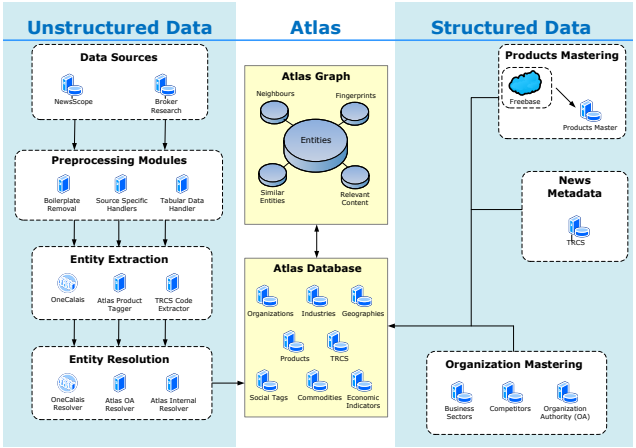


Figure 2: Graph Construction

4.2 Entity Extraction and Resolution

Once pre-processing is complete, we aim to find the entities mentioned in each document, and where possible, resolve these mentions to a structured dictionary of known entities, forming the basis of the Atlas graph. As we discuss in the next section, attempting to deal with every instance of every entity type supported by Atlas that was mentioned in five years of news turned out to be a significant challenge.

Entity extraction was mostly dealt with through OneCalais, the enterprise version of OpenCalais[6] provided by ClearForest, a Thomson Reuters company. OneCalais uses NLP and machine learning techniques and has been commonly used by the research community [26, 17, 16]. Whilst Atlas contains additional features to extract entities that Calais currently has little support for, these are quite trivial in nature and focused mainly on basic string matching. The Calais platform also provides some degree of entity resolution against structured data, but in cases where it was unable to do so, the job of entity resolution fell to Atlas. Atlas' entity resolution mechanisms included string similarity, aliases from user feedback and expert-supplied rules.

In addition to extracting entities from documents, OneCalais also provided a list of *social tags* associated with each document. Social tags are a new feature within OneCalais that provide automated classification based on an organic taxonomy derived from Wikipedia.

4.2.1 General Resolution Issues & Challenges

It was discovered that in our structured datasets, entities are often listed by unintuitive names, such as the full legal names of organizations. Large corporations may have many subsidiary companies, and structured data on the relationship between these subsidiaries was not always available. For example, Microsoft has at least 70 separate legally-recognized organizations worldwide. Where a parent-child chain was available in structured data, Atlas made use of it by counting each mention of a child as a mention of itself, and of every parent in the chain. A configurable parameter allows Atlas users to select the degree of desired granularity, i.e., from considering every legal entity to only considering top-level parent organizations. In cases where this information was not available in structured data, a number of heuristics were applied in an attempt to identify such relationships. Of particular utility was a list of legal suffixes for companies that was made available to us by domain experts.

[6]http://www.opencalais.com/

Entity extraction is not an exact science and a number of erroneous matches were returned by the OneCalais engine. Since our significance computation method tends to assign low scores to connections involving erroneous entities, these had little impact on aggregate results, but they became prominent when we developed user interfaces to display emerging or unusual relationships. Some common patterns that required special handling involves the journalistic style of listing companies in news, e.g. "Intel Samsung Toshiba join hands to halve chip size — Nikkei" was originally extracted as a company called Intel Samsung Toshiba.

Certain companies with short names containing common English words proved very difficult to accurately extract and resolve to. Examples include Business (a French publication) and General Corporation (an American realtor). Where non-English words were present in the dataset, these often provided entity resolution problems, a common issue being person names extracted as organizations.

A final common issue in entity resolution was that of distinguishing between mentions of a location, and mentions of that location's governing body: 'New York State,' for example, is interchangeably used to refer to either the geographical area or the governing body of that area.

4.2.2 Media Companies

It is an unfortunate fact that those who report the news are often incidentally mentioned within it, despite having no connection to the story itself. Whilst our boilerplate removal algorithm was able to strip out the vast majority of inappropriate mentions of media companies, a significant number remained in free text. An example from our dataset is "...from Morgan Stanley told Reuters in an interview on Friday that." Whilst entity extraction and resolution was perfect in this example, it skewed Atlas' results as no useful connection exists between Morgan Stanley and Reuters. Ratings agencies were also affected by this problem.

The current version of Atlas does not handle these situations well. In the future, we plan to incorporate NLP-based rules to handle common patterns that exhibit this problem, and also to utilize relevance scores returned by OneCalais.

4.2.3 Financial Institutions

It can be seen from a cursory examination of our data that large financial institutions truly have a sphere of influence that touches nearly every point of the corporate world. Whether providing funding, services or advice, or buying or selling products, genuine relationships seem to exist between any given large financial organization and tens of thousands of other companies. Whilst aggregate results again solve the issue of ranking a large financial's relationships, two problems emerge from this situation.

The first is that, for small companies with few mentions, a large financial may be seen not only as a relatively strong connection, but also as a similar entity (given that the two appear in the same documents with relatively high frequency and therefore acquire the same neighbor entities). The second is that it can become very difficult to statistically guess or use attributes about the large financial: whilst they may be based in a given geography for example, it may be that the companies they are advising are doing business in a completely different geography. The chained relationship which is therefore formed is not always true or meaningful, and can cause issues with a number of metrics.

4.2.4 Optimization

A final issue in dealing with large datasets was that of optimization. As an example, there were over 0.75 billion pairs of organizations for which the similarity scores needed to be calculated in the graph evaluated in this paper. Distributing the workload achieved only so much, and the use of vertical bitmaps to compare entities (e.g. based on document co-occurrence or possession of structured attributes) was a key factor in improving the performance of some of the algorithms used to generate the final graph.

4.3 Significance Scores

As we have discussed in Section 3.2, Atlas uses entity co-occurrences in unstructured documents as primary means to establish connections between entities. This approach indeed maximizes the recall, but results in a lot of noisy connections. We address this problem by assigning a significance score to each edge, where higher values indicate stronger connections. Since we use a directed graph, there are always two edges between each pair of connected entities, each of which is assigned a different significance score. This allows Atlas to model a common real-life situation where a given entity E_1 may be very significant for a connected entity E_2, but E_2 may not be equally significant for E_1. Consider, for example, two companies that compete in an area that represents the core business for one company but only a small fraction of the other company's business.

Figure 3 presents actual Atlas significance scores between Facebook and some of its neighbors. The scores indicate that Twitter is more significant for Facebook than Microsoft, and both are more significant than Apple. In contrast, Facebook is important for Microsoft but not equally significant: Microsoft is a major shareholder and investor in Facebook, but is also involved in a variety of other business areas. The relationship between Facebook and Google exhibits a similar behavior. By contrast, Facebook is the major threat to MySpace's core business, making the most significant connection in Figure 3 the MySpace —> Facebook edge; whereas MySpace's actions are now much less important to Facebook, which is clear from the fact that the connection in the opposite direction is of nearly $\frac{1}{3}$ the strength.

To compute the significance scores, we have considered a variety of strategies, and finally selected a few that capture different aspects of the relationship and produce superior results when combined (Section 6.1). Therefore, the Atlas significance scores are computed as a weighted average of multiple factors (where weights were empirically selected). The factors we have evaluated (Section 6.1) include:

a) **Interestingness:** Computed by considering the source and target entities as two variables, populating a 2x2 contingency-table with their frequencies from all available documents (as in Section 1 of [29]), and applying an interestingness measure on the contingency table. Originally proposed for finding interesting association rules, interestingness measures have been successfully used in many applications [20]. We evaluated all measures in Table 5 of [29], and selected "Mutual Information" based on its superior performance on our data (another measure "Added Value" was very close in terms of performance). We have omitted the interestingness measure comparison for the reason of space.

b) **Recent Interestingness:** Same as interestingness, but

computed only using documents in a user-definable (fixed to 6 months in our experiments) recent period. This factor aims to boost emerging relationships.

c) **Validation:** A value of 1 if the relationship between source and target entities was validated by a human expert (in available structured data), 0 otherwise.

d) **Common Neighbors:** A percentage of the degree-1 neighbors of the source entity that also occur in the degree-1 neighborhood of the target entity.

e) **Industry Overlap:** A percentage of the industries in source entity's neighborhood that also occur in the degree-1 neighborhood of the target entity.

f) **Geography Overlap:** A percentage of the geographies in source entity's neighborhood that also occur in the degree-1 neighborhood of the target entity.

g) **Temporal Significance:** A comparison of the recent interestingness value, with an interestingness value computed from historic (non-recent) documents giving a value of 1 to the factor if the former increased by more than α, a value of 0 if the the recent interestingness decreased by more than α or a value of 0.5 otherwise, where α was fixed to 0.05. This factor rewards relationships that had gained strength in the recent time period, and penalizes relationships that had lost strength.

h) **Element of Surprise:** Using the same definitions of recent and historic documents as in the previous factor, this factor is assigned a value of 1 if the source entity's neighborhood contains any new industries or geographies in the recent period that did not occur in the historic period, and the target entity shares at-least one such industry or geography, 0 otherwise.

4.4 Entity Fingerprints

A primary business use-case from our organization for Atlas was the need to find companies that were *similar* to a given company, overlooking any geographical differences. Should such functionality be made available, a party with an interest in a successful catering company in Denmark would, for example, be able to find similar opportunities for investment in the emerging Chinese markets. Once geographical boundaries were crossed, it soon became obvious that two similar companies – in our example, companies that may be affected by the same commodities and laws and share similar types of neighbors – may never co-occur in documents, and would therefore never have an edge between them in the Atlas graph. However, the aggregate properties of each company's neighbors, and the other information we were able to infer about them from Atlas, allowed us to construct a *fingerprint*, an abstract representation of the company which could be compared to any other fingerprint in order to calculate a *similarity score* and support this use-case.

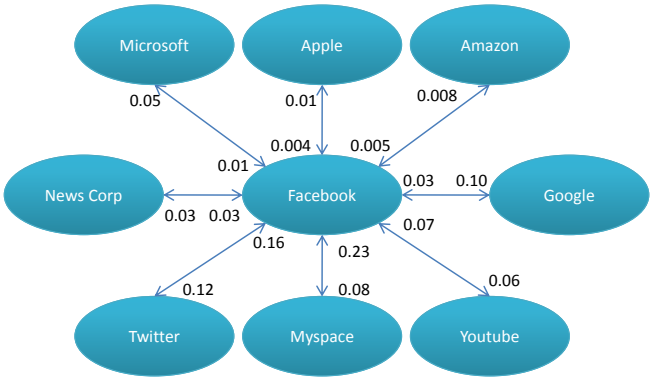


Figure 3: Atlas significance scores between Facebook and some of its neighbors

An entity's fingerprint is a multi-dimensional abstraction of the entity based on a number of its attributes. Once a fingerprint is obtained for an entity, it may then be compared to fingerprints of other entities to understand the similarities and differences that exist. Fingerprints may therefore also be used to generate feature vectors in classification and clustering tasks. In the version of the Atlas graph presented in this paper, fingerprints were calculated and used only for organizations, to support the business case described above.

The factors used to generate fingerprints were, in the majority, related to the neighborhood of a given organization: the entities of a given type to which it was related and temporal aspects of these (e.g. one attribute was 'emerging neighbors', thus fingerprints can represent a point-in-time view of an entity). Structured information about the organizations, such as the country they are incorporated in and the industries they are known to operate in, were also incorporated into the fingerprints. Finally, we have used entity to document relationships to add the top-k classification codes (i.e., TRCS codes) and top-k social tags (i.e., Wikipedia article titles related to a document, as determined by OneCalais) for each company to the fingerprint. This was achieved by sorting each TRCS code or Social Tag with respect to the number of documents that contained the target company and were also assigned with the TRCS code or Social Tag, and selecting k most frequent results.

Each company's fingerprint therefore includes the following attribute groups: industry hierarchy; geography hierarchy; related industries; related geographies; related macroeconomic indicators; related commodities; related TRCS codes; related social tags; and related entities (in two groupings — those that are emerging and those that are stable).

A feature vector is created for each attribute group, and a similarity score between two fingerprints is computed as a weighted linear combination of the cosine similarity scores of group feature vectors, where weights were empirically selected by domain experts. This score represents how alike two companies are, rather than simply how connected they are in the graph.

4.5 Characteristics of the final graph

Presented in Table 1 are the number of distinct entities (nodes) – a total of 85,163 – found in the graph. Between these, 13.3 million connections (edges) were found, with each node having an average of 156 connections (std. dev. 886). On a single server with 32 CPU cores and 128GB of RAM, the graph took a little over 9 hours to build, and included the creation of data-stores such that the vast majority of queries were comfortably sub-second.

The concept of validation is important in a number of calculations in the Atlas system, and in displaying information back to end users. Furthermore, in the case of organizations, non-validated nodes were almost always entity extraction or entity resolution errors. In use-cases that reward surprise, such as those looking for emerging trends or uncommon connections, these caused a significant issue.

An organization is said to be validated if it is present in

Type	Count	% Validated
Organization	44,117	95%
Geography	37,902	3%
Product	2,762	19%
Industry	189	100%
Commodity	130	100%
Macroeconomic	63	100%
Total	85,163	52%

Table 1: Distinct Entities in the Graph

a database containing all public, and a good number of private, companies. This database is maintained and updated by a large team of analysts and researchers and powers some of our company's largest commercial products. It took the Atlas team hundreds of man-hours to reach 95% validation level for organizations, and this work led to improvements to both the internal database and the OneCalais core engine.

Geographies were also validated against an internal database, but in this case the 3% validation figure isn't quite as damning. Whilst a significant number (nearly 60%) of geographies in the graph are invalid, these occur infrequently and do not affect many of the primary use-cases of Atlas. The remaining number are simply small or hyper-localized geographies not meriting a place in the database used, but extracted nonetheless by the OneCalais engine.

Products were validated against a database compiled specifically for this project from the open-source repository Freebase. Those that are not validated were returned from the OneCalais entity extraction module and appear to be valid, in the majority, although no formal evaluation of this has yet been performed.

5. THE USER INTERFACES

In order to effectively explore a complex graph with tens of thousands of nodes and millions of edges, a set of novel user interfaces were designed and developed. The focus was to deliver specific business use-cases which would directly benefit the customers of Atlas. A select few rich user interfaces were developed in Adobe Flex & Flash with a larger number of HTML "helper pages" made available for those who wished to drill deeper into the raw data. Examples of these include the ability to look up a filtered list of any entity's connections, together with supporting documents; the ability to view a side-by-side overlap of the characteristics of two similar entities; or the ability to provide manual feedback for future resolution (e.g. that 'MSFT' should be resolved to 'Microsoft' where no other evidence is available).

5.1 Find Similar

The Find Similar view in Atlas (Figure 4) makes use of the fingerprint (see Section 4.4) of an organization. For any given company, the Find Similar view shows a list of the most similar companies along with their fingerprints in a given time range, and highlights the reasons for the match.

Other useful information returned includes an overview of the fingerprint for all companies returned, and a graph showing how each resulting company has trended with the query company over a recent period, with document links.

6. EMPIRICAL EVALUATION

The output of machine learning and information retrieval systems is typically evaluated against human-labeled gold data, using standard objective metrics such as precision and recall. There are an unknown number of useful 'questions' that could be asked of a system like Atlas, and many of these have very subjective and constantly-changing answers, making it extremely difficult to obtain and maintain labeled data for evaluation. Even for well-defined tasks (e.g., retrieving top-K neighbors for a given entity), a comprehensive evaluation was resource-prohibitive because of the number of entities available in the system.

Therefore, we have evaluated Atlas by selecting representative samples of entities; obtaining Atlas output for two common tasks; and having domain experts review and grade the output. The next two sections present the results of these experiments. We also present a subjective comparison of Atlas output against existing systems and analyze the fingerprint overlap between two entities.

6.1 Significance Scores

Fifty companies were randomly selected for evaluation from the Justmeans Global 1000 Sustainable Performance Leaders list[7], twenty-five from the top 500 and twenty-five from companies ranked between positions 500 and 1000. For each of these companies, the most significant fifty connections in Atlas were manually evaluated, with human experts identifying each result as either a genuine real-world connection; as a connection with no meaningful foundation in the real-world; or as a duplicate connection (including a duplicate of the source company).

6.1.1 *Significance Factors*

A number of significance factors were discussed in Section 4.3, but each of these was found to incorrectly skew the results when presented with certain classes of entity, or types of source information. A weighted approach, was therefore taken, and its precision evaluated relative to each individual factor. A precision of 100% would be achieved where a significance factor's top-K connections were all valid, non-duplicate connections, where K is the total number of valid, non-duplicate connections that exist for the given source company. The graphs in Figure 5 present the results of this experiment, with the y-axis representing precision and the x-axis enumerating each of the fifty source companies considered, sorted in decreasing order of weighted significance precision.

[7]http://www.justmeans.com/top-global-1000-companies

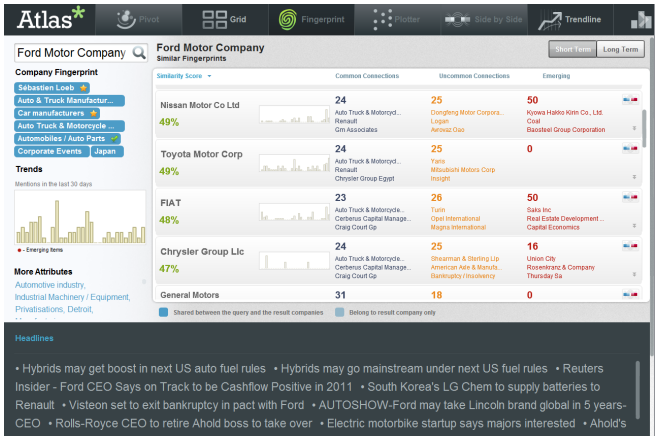


Figure 4: The Find Similar View

Metric	Average Precision
Weighted Significance	58.69%
Recent Interestingness	55.57%
Temporal Significance	54.08%
Interestingness	52.4%
Validation	50.49%
Common Neighbours	20.6%
Geography Overlap	17.43%
Industry Overlap	14.83%
Element of Surprise	8.06%

Table 2: Average Performance of Significance Factors

Whilst individual factors occasionally achieve higher precision than the final weighted significance, the average performance of the latter was found to consistently outperform any individual factor (see Table 2).

6.1.2 *Top Neighbors*

In the previous section, we have evaluated various significance factors and concluded that our weighted significance score outperforms individual factors. We now evaluate its usefulness in a real scenario. A major business use-case for our organization was that of being able to identify the important commodities for any given geography — currently a manual and error-prone task. Twenty countries of varying size were selected for evaluation, and their top-ten commodity neighbors in Atlas were evaluated by a domain expert.

In Figure 6 we present the results of this evaluation, sorted in the increasing order of the number of documents that contained each country in our dataset. We note that without any thresholding on the Atlas results, the system was still able to achieve 76% precision, and also that the majority of false-positives found were located towards the lower ends of Atlas ranking lists. A significant exception here was Israel's most significant commodity, Nuclear Power: an unfortunate example of entity resolution gone awry: the supporting documents were referring to nuclear weapons.

6.2 Fingerprint Factors

6.2.1 *Similarity*

It is difficult to find ground truth against which to evaluate the lists of most similar companies produced by Atlas for a given entity. *Similarity* is a vague and subjective term, and often closely coupled with the context in which a user is searching. Here, the Atlas system is compared against two well-known and well-used datasets, Reuters Knowledge

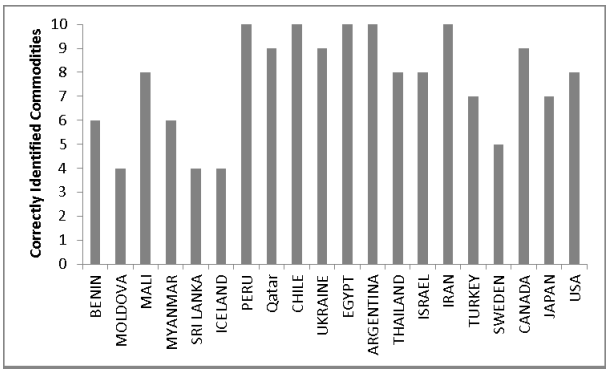


Figure 6: Top Commodity Neighbors for Countries

Company	Atlas	Google	Reuters
Atlas Top Ten			
General Mills	1	0	0
Keebler Foods	2	0	0
Ralcorp Holding Inc	3	9	0
Green Mountain Coffee	4	0	0
Tate Limited	5	0	0
J&J Snack Foods Corp	6	7	0
Campbell Soup Company	7	3	3
J.M. Smucker Company	8	0	0
Koninklijke Wessanen Nv	9	0	0
Premier Foods	10	0	0
Google Top Ten			
Tasty Baking Company	0	1	0
Kellogg Company	16	2	0
Grupo Bimbo S.A.B.	20	4	2
Sara Lee Corp	12	5	1
Vitafort Intl. Corp	0	6	0
Treehouse Foods Inc	14	8	0
Snyder S Lance Inc	17	10	0

Table 3: Similar Companies for Flowers Foods Inc

and Google Finance's Related Companies. Reuters Knowledge provides a manually-maintained list of similar companies generated by analysts from company publications and releases. Google, like Atlas, uses an algorithmic approach.

The "most recent quarter" option was selected in Google Finance, and a three-month time period selected in Atlas. Reuters Knowledge offers no time-sensitive options, and may list between 0 and 30 similar companies. Google offers only the top-ten most similar companies (and in the majority of cases seen, returns 10 companies), whereas we have access to the entire list for Atlas, which, without thresholding, often contains in excess of 30 results. The company we selected for detailed evaluation, based on the belief that it represents some of the key positive and negative features of Atlas in evaluating similar companies compared to Google Finance and Reuters Knowledge, is Flowers Foods Inc, a company based in the USA that markets a range of top-selling consumer brand packaged foods. Table 3 shows the top similar results for Flowers Foods in each dataset.

The first point to note is that entity resolution is ever an issue when dealing with organizations. Whilst Keebler Foods and Keebler Holding are indeed separate legal entities, for the vast majority of use-cases, the distinction is not a useful one. Further, it is not rare for companies to merge or be acquired, and the multiple processes by which this may occur may also lead to confusion in a system.

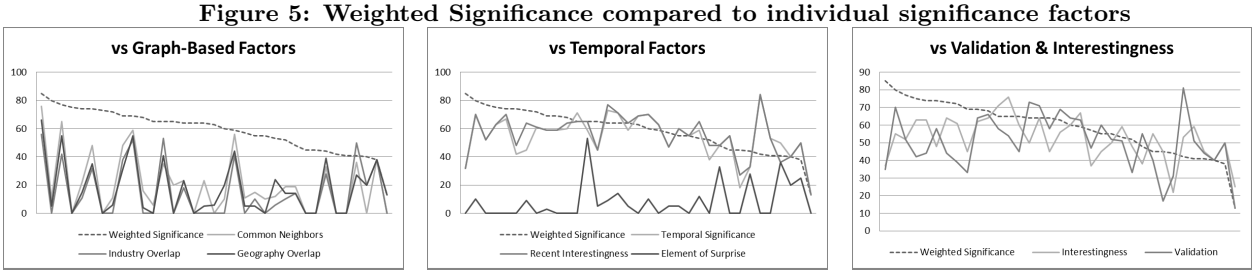
Whilst both Atlas and Google were able to nearly flawlessly[8] extract companies with very similar business offerings as Flowers Foods Inc, their ranking differed. Only three of the Atlas top 10 were found in the Google top 10, all the Google top 10 except Vitafort were found within Atlas' top 20. The only clear pattern was that Atlas seemed to favor similarity based on product offerings (core business), whereas Google favored companies based in the same country – the USA – with the exception of Grupo Bimbo S.A.B., based in Mexico (with a significant American presence).

Over a sample of 25 companies from various industries and geographies, 87% of Google top 10 similar companies appeared in Atlas' top 20, and 34% of Atlas' top 10 similar companies appear in Google's top 10.

Reuters Knowledge returned a mere three similar companies for Flowers Foods, all of which were also identified by

[8]Vitafort being the exception, it appears to be an erroneous match from Google Finance: Vitafort is wholly involved in entertainment intellectual property.

Figure 5: Weighted Significance compared to individual significance factors



both Atlas and Google. There was a strong correlation here between the top companies in Google and the top companies in Reuters Knowledge, whereas Atlas ranked the three companies from Reuters in positions 12, 20 and 7 respectively — suggesting that Atlas' ranking methodology may differ to that used in current products. Of the 25 sample companies, Flowers Foods was found to be one of the worst examples of Atlas' rankings corresponding with Reuters Knowledge. When querying on Toll Brothers, Inc, for example, Reuters Knowledge provided five results which Atlas ranked in positions 3,9,5,1 and 2 respectively. For a larger company, Google, Reuters Knowledge provided seven results, which Atlas ranked in positions 1,2,10,19,14,4 and 6 respectively.

We stress once again that Atlas is not intended to reproduce the results of either Google Finance or Reuters Knowledge, but find that correlation with these two well-used tools, together with expert opinions provided to us, show the utility of Atlas-generated similar companies.

6.2.2 Overlap

In addition to using entity fingerprints to compute similarity, it may often be useful to understand in detail the overlap between two entities. Take, for example, YouTube (now owned by Google) and Facebook. These are both prominent Web companies, which shows in the overlapping attributes they have. However, their differences may also be easily seen at a glance using the companies' fingerprints (Table 4), in which a score of -1 implies an attribute owned only by YouTube, and a score of 1 an attribute owned only by Facebook. A score of 0 implies a perfect overlap.

$$Overlap(a,b) = \begin{cases} \frac{b}{a} - 1 & a \geq b \\ 1 - \frac{a}{b} & a < b \end{cases} \quad (1)$$

In this example, we can see that the Atlas graph correctly provides a strong bias towards YouTube in considering the geography attribute "San Mateo ," YouTube's home county. Despite having no presence in San Mateo, Facebook is somewhat linked to it by documents mentioning both YouTube (with San Mateo as a corollary) and Facebook, a common problem in evaluating similarity for two entities that are connected in the graph. An extraction error resulting from the suffix -ville has led to the popular Facebook social game "Farmville" being identified as a geography, and this is therefore understandably considered related only to Facebook; it will provide negative evidence in computing a similarity score between these two companies.

In the Social Tags attribute type, "World Wide Web" is correctly a strongly shared connection, providing positive evidence towards similarity, whereas "Video Hosting" for YouTube and "Social Network Service" for Facebook exemplify the core differences between these two companies' activities on the Web.

Geographies	
San Mateo	−0.81
Farmville	1
Social Tags	
World Wide Web	0.07
Video Hosting	−1
Social Network Service	1
Recent Organizations	
Viacom Inc	−0.77
Twitter	0.42

Table 4: Fingerprint overlap for YouTube and Facebook

In the Recent Organizations attribute type, Viacom Inc was an attribute most related to YouTube, and Twitter an attribute biased somewhat towards Facebook. Viacom offers many video services, and has a natural relationship with YouTube, but also made offers to buy Facebook on multiple occasions. Twitter competes with Facebook for the real-time microblogging market: although it is mentioned in many news stories alongside YouTube, the relationship is often incidental. The final bi-directional similarity score between YouTube and Facebook given was 56%, and each of these companies was top of the others' most similar list. Facebook's second most similar company (51%) was LinkedIn, which offers professional social networking; YouTube's second most similar company (49%) was its parent, Google.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented Atlas, an entity-centric information analysis, content aggregation, and visualization platform that uses all available structured and unstructured content to automatically model the ecosystem around entities. Atlas uses a novel method to compute the significance of entity connections, and automatically summarizes key entity attributes into a fingerprint. The results of our objective and subjective evaluation indicate that our significance computation method is able to reward important and interesting connections, and the fingerprints are useful in finding similar companies, an important use case in our domain.

In the future, we plan to work on automatically identifying relationship types for entity connections, incorporating Web and markets fundamental data in graph generation, and improving entity resolution for geographies and products.

8. REFERENCES

[1] E. Agichtein and V. Ganti. Mining reference tables for automatic text segmentation. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 20–29, New York, NY, USA, 2004. ACM.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165, 2009. The Web of Data.

[3] V. R. Borkar, K. Deshmukh, and S. Sarawagi. Automatic text segmentation for extracting structured records. In *International Conference on Management of Data*, 2001.

[4] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proceedings of the 6th workshop in very large corpora*

[5] P. Bouquet, H. Stoermer, C. Niederee, and A. Maña. Entity name system: The back-bone of an open and scalable web of data. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing*

[6] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*. The Association for Computer Linguistics, 2006.

[7] S. Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 571–580, New York, NY, USA, 2007. ACM.

[8] S. Chakrabarti, J. Mirchandani, and A. Nandi. Spin: searching personal information networks. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 674–674, New York, NY, USA, 2005. ACM.

[9] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *In Proc. 2007 Joint Conference on EMNLP and CNLL*, 2007.

[10] A. Culotta, T. Kristjansson, A. Mccallum, and P. Viola. Abstract corrective feedback and persistent learning for information extraction. *Artificial Intelligence*, 170, October 2006.

[11] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[12] T. G. Dietterich. Machine learning for sequential data: A review. In *Structural, Syntactic, and Statistical Pattern Recognition*. Springer-Verlag, 2002.

[13] D. Ferrucci and A. Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10:327–348, September 2004.

[14] J. Hassell, B. Aleman-Meza, and I. Arpinar. Ontology-driven automatic entity disambiguation in unstructured text. In *The Semantic Web - ISWC 2006*, volume 4273 of *Lecture Notes in Computer Science*, pages 44–57. Springer, 2006.

[15] J. R. Hobbs, J. Bear, D. Israel, and M. Tyson. Fastus: A finite-state processor for information extraction from real-world text. pages 1172–1178, 1993.

[16] F. Hopfgartner and J. Jose. Semantic user modelling for personal news video retrieval. In S. Boll, Q. Tian, L. Zhang, Z. Zhang, and Y.-P. Chen, editors, *Advances in Multimedia Modeling*, volume 5916 of *Lecture Notes in Computer Science*, pages 336–346. Springer Berlin / Heidelberg, 2010.

[17] F. Iacobelli, L. Birnbaum, and K. J. Hammond. Tell me more, not just "more of the same". In *Proceedings of the 15th international conference on Intelligent user interfaces*, IUI '10, New York, NY, USA, 2010. ACM.

[18] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01*, pages 282–289, 2001.

[19] M. D. Lieberman, S. Taheri, w. Guo, F. Mirrashed, I. Yahav, A. Aris, and B. Shneiderman. Visual exploration across biomedical databases. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, March 2011.

[20] H. H. Malik, J. R. Kender, D. Fradkin, and F. Moerchen. Hierarchical document clustering using local patterns. *Data Min. Knowl. Discov.*, 21:153–185, July 2010.

[21] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *In Proceedings of KDD ï£¡05*, pages 198–207. ACM Press, 2005.

[22] E. Minkov and W. W. Cohen. Improving graph-walk-based similarity with reranking: Case studies for personal information management. *ACM Trans. Inf. Syst.*, 29:4:1–4:52, December 2010.

[23] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 446–453, New York, NY, USA, 2004. ACM.

[24] E. Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the eleventh national conference on Artificial intelligence*, AAAI'93, pages 811–816. AAAI Press, 1993.

[25] S. Sarawagi. Information extraction. *Found. Trends databases*, 1:261–377, March 2008.

[26] S. Sen, N. Stojanovic, and R. Lin. A graphical editor for complex event pattern generation. In *Proceedings of the Third ACM International Conference on Distributed Event-Based Systems*, DEBS '09, New York, NY, USA, 2009. ACM.

[27] D. Shahaf and C. Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, New York, NY, USA, 2010. ACM.

[28] K. Takeuchi and N. Collier. Use of support vector machines in extended named entity recognition. pages 119–125, 2002.

[29] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, New York, NY, USA, 2002. ACM.