ABSTRACT
Gathering data from open sources about the activities and events related to competitor organizations is a vital task. Such data be used to understand the business strategies, plans and models of the competitors. This in turn can help an organization to successfully adapt and evolve in a fast-changing market environment. In this paper, we describe an Information Extraction tool called INX that we have developed and its use to gather competitor intelligence by analyzing news items. We describe INX’s innovative multi-pronged approach to information extraction. We provide examples of information extracted by INX from news items about financial events like mergers and acquisitions as well as award of projects and contracts.

1. INTRODUCTION
To be at the forefront in today’s competitive world, it is important to have up-to-date information about your competitor organizations. Vital information about the activities as well as strengths and weaknesses of the competitors can help an organization in adjusting business strategies to gain an edge over the competitors. This is the task of gathering competitor intelligence. This task is critical for any modern business because the business and market for products and services change rapidly.

Competitor organizations are continuously involved in many financial events: mergers and acquisitions, awards of projects, deals and contracts, building and offering new products or services, launching of marketing campaigns and sales initiatives (such as price cuts, discounts and loyalty rewards), opening of offices in new territories, announcements of alliances and partnerships with other industries, announcements of technology development programmes with academic institutes, announcement of financial results, joining and resignation of influential people and so on.

Many of these events are reported in various public for and media. The main question is: how does an organization obtain information about its competitors’ activities using open sources? The business news stories and articles published in financial newspapers, magazines and on various web sites are a good source, if used effectively. Unfortunately, given today’s information explosion, manually analyzing all possible news sources is not viable due to numbers and size of the news items generated by the media or available on web sites. Such a manual process is time consuming, repetitive (boring) and inconsistent as well; the quality of manually extracted information may well vary with the expertise and experience of the persons doing the work.

The goal then is to automatically analyze a given set of news items and gather factual data about the financial activities and events related to specific competitor organizations.

The financial events of interest are of many different types, as listed above. Ideally, the system should analyze the given collection of news items and create a structured repository (e.g., relational tables), one for each type of financial event, and that should fill a table up with details of the corresponding financial events. Data in these tables can then be analyzed using statistical or data mining techniques (e.g., to build models of a competitor’s business strategies).

As an example, Fig. 1 shows two news items, each of which reports a company acquiring (purchasing) another company. An automatic information extraction program should analyze these 2 news items and produce two records (one for each acquisition event) in a relational table for M&A events, as shown below (this is actual output of our INX tool). Further details about the event can also be extracted from the news item and added to the table as extra columns; e.g., % stake of the acquirer company, location and nature of business of the acquired company etc.

<table>
<thead>
<tr>
<th>Acquired by</th>
<th>Company acquired</th>
<th>Acquisition cost</th>
<th>Acquired co. turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zensar Technologies</td>
<td>ThoughtDigital</td>
<td>--</td>
<td>27 million dollars</td>
</tr>
</tbody>
</table>
This paper is organized as follows. In section 2, we describe the general problem of extracting structured information from text and then describe an information extraction engine INX built at Tata Research Development and Design Centre (TRDDC), Pune. We also describe the technology within INX that performs information extraction. In section 3, we then describe how INX was used to automatically analyze a set of financial news items to gather competitor intelligence. In section 4, we present our conclusions and discuss further work.

News 1:
ZENSAR ACQUIRES NEW YORK-BASED THOUGHTDIGITAL. Zensar Technologies has completed the acquisition of ThoughtDigital, a New York-based software firm that specializes in Oracle applications. ThoughtDigital, owned by SOA Software, is a profitable East Coast (USA) systems integrator with a turnover of 27 million dollars for the year ended December 31, 2006. It employs 120 Oracle consultants and has a strong client base in verticals such as communications and media, financial services, consumer products and services. Zensar expects to achieve great synergies with ThoughtDigital in the Oracle space, given the rapid growth of its Enterprise Application Solution business in the US market.

News 2:
IT solutions provider Mastek Ltd on Thursday announced the acquisition of US-based Vector Insurance Services LLC (Vector) for $9 million. Vector is a technology solutions provider and third party administrator that focus on the North American life and annuity insurance industry, having two of America’s largest insurance carriers as its customers. The acquisition will be done by Mastek’s wholly-owned US subsidiary MajescoMastek which will be acquiring a 90 percent equity stake in Vector in an all-cash transaction for $4.5 million, its Chairman and Managing Director, Sudhakar Ram, told reporters here. This acquisition is revenue and earnings accretive to Mastek and is expected to strengthen the opportunity pipeline for the company in the insurance

2. INFORMATION EXTRACTION (IE)

2.1 Problem Definition and Applications
Most organizations and individuals now have access to vast repositories of documents. Vital information and knowledge is buried deep within these documents. There is a clear emerging need in large enterprises for software systems that can automatically analyze these documents and extract the hidden information from them. The extracted information can be kept in databases, which can then be shared among the employees. These databases can be analyzed using statistical and data mining techniques. The knowledge gained can be used for more informed decision making.

Information extraction (IE) [3] is an emerging field within the broad area of text mining that attempts to automatically analyze a given set of text documents and produce a structured repository (e.g., a relational table) containing information about specific events or activities in those documents (Fig. 2). The extraction is based on a set of user-defined patterns, which are also input to the IE system. IE techniques are based on those in the neighbouring technical fields of natural language processing, statistical language modeling, data mining, machine learning and statistics.

Figure 1. Two sample news items reporting mergers and acquisitions.

Figure 2. Schematic for an Information Extraction tool.

There is a need for specialized IE techniques, because it is difficult to directly analyze text – e.g., using pattern matching or statistical / data mining techniques. There are several difficulties due to the special nature of the text.

1. Unlike data in a table, text is unstructured i.e., does not have fixed columns (fields).
2. Unlike data, text cannot be manipulated using simple operators (e.g., +, -, *, /). You cannot take an average of a sentence, for example!
3. Unlike data, text has a complex semantics i.e., meaning, which needs to be understood before we can extract useful facts from text. “Windows crashed” means something special in a software system’s manual; it does not refer to a problem in a wall aperture. Further, sentences in a text can have implicit connections; e.g., the pronoun “he” in a sentence may refer to somebody mentioned in one of the earlier sentences. Correct anaphora resolution requires complex analysis of the text.
4. Unlike data, text is inherently ambiguous. A value of 20,000 in the salary column of an employee table has a specific meaning. However, the same fact may be stated in many different ways in text. For example, the sentences “your service is slow” and “your administrative efficiency is low” usually mean the same thing, though they use very different words.
5. Data usually contain only a few errors, if any; e.g., due to data entry errors. Even measurement data from instruments contains well-restricted (e.g., Gaussian) noise. Text, on the other hand, may contain much noise and errors; e.g., spelling and grammatical mistakes, missing punctuations, use of acronyms etc.
• Extract skills and experience from resumes
• Automatically extract project details from tender notices in newspapers
• Extract maintenance knowledge (e.g., problem, symptoms, causes, solution) from customer support and equipment maintenance document
• Extracting details of important events from news items
  – Financial events “Mr. Shah left IBM and joined TCS as VP”
  – Terrorist attacks “3 killed in a bomb blast in a hotel in Jammu”
  – Criminal activities “Dacoits struck at a farmhouse near Pune”
• Extract complaints and specific actionable suggestions from customer (or employee) satisfaction surveys
• Extract accident and injury details from auto insurance claim forms

2.2 Techniques for Information Extraction

It may appear that one can write simple regular expressions (e.g., in Perl or C#) and extract the relevant facts from each news item. For example, the following Perl regular expression

\[(d+).*?BHK.*?d+).*?sqft.*?rent.*?d+\]

correctly extracts the number of bedrooms (3), size in sq. feet (1640), and rent amount (20000) from the following sentence in a classified advertisement:

3BHK 1640 sqft fully furnished at Parmar Garden wanjowrie rent 20000/-."

However, the regular expression would fail if the number of bedrooms or size were mentioned differently (“3 bedrooms” or “165 sq. meters”). It may appear that we can patch up the regular expression to cover such special cases. While such literal matching may work to some extent (though it will never cover all possible flat-for-rent advertisements!), we run into real difficulties if we want to extract locality of the flat. It is difficult to recognize that a group of words (“Parmar Garden wanjowrie”) forms a location; it needs a deeper analysis of the text.

The approach we have taken is as follows (Fig. 3). We apply several natural language processing techniques to the given text to perform different types of analysis. We then combine the results of all these analyses and embed these results in the original text itself to obtain an enriched representation of the original text. We then use the standard regular expressions designed to extract the desired information from the enriched text.

The basic advantage of this approach is as follows. The enriched text representation explicitly contains a lot more information than the original text. This allows us to write higher level (i.e., more abstract) regular expression that match the results of analysis rather than literal match within the text.

The parser analyzes each sentence and identifies several syntactic constituents (e.g., noun phrases, verb phrases, preposition phrases, clauses etc.) within the sentence and their hierarchical and positional relationships among each other. These relationships are displayed in the form of a parse tree.

The semantic role analysis tool identifies the various semantic roles for the given sentence. Semantic roles (also called thematic roles) are the various semantic functions performed by the constituents of a sentence. In the sentence “John broke the window with the hammer”, John is the AGENT (who performs the action described), the window is the THEME (the thing affected by the action described), the hammer is the INSTRUMENT (the tool, material or force used to perform the described action).

There are many other semantic roles such as BENEFICIARY, TO-POSS, AT-POSS, FROM-POSS, TO-LOC, FROM-LOC, AT-TIME, FROM-TIME, AT-VALUE, FROM-VALUE, CO-AGENT etc. Semantic roles are related to but distinct from grammatical relations such as SUBJECT, OBJECT etc. Semantic roles are identified using their meaning, not their syntactic label or form. Most semantic role analysis techniques related to but distinct from grammatical relations such as SUBJ

Fig. 4 shows an example of a sentence and its enriched representation in the form of an enriched parse tree. NP, VP etc. indicate the syntactic categories. Tag ORG indicates that the words (and corresponding subtree refer to an organization e.g., “The Bahraini bank” and “Infosys”. The semantic roles K1 and
K2 respectively indicate the SUBJECT and OBJECT of the verb “selected”; e.g., “The Bahraini bank” and “Infosys banking solution”. Fig. 5 shows the corresponding enriched text obtained by “flattening” the enriched parse tree. The start and end positions of syntactic categories, semantic roles and named entities within the text are indicated by XML-like tags.

Fig. 4. Enriched parse tree for an example sentence.

The Bahraini Bank has selected Infosys banking solution Finacle.

Fig. 5. Enriched text representation for an example sentence.

The enriched representation of the original text is searched using the usual patterns of Finance Analysis. However, because of the rich information added to the text, the patterns can now be much more complex and abstract (at a high conceptual level). Some examples of such patterns are given below. For each pattern we also show an example sentence and underline the parts extracted by the pattern. The patterns are stated in plain English (rather than as regular expressions) for easy understanding.

1. (Projects awarded): Find sentences containing verbs {award} and noun {project, system, solution} and return the subject and object of the verb if both are named entities of the type ORGANIZATION.

Example: Ferrari has awarded TCS a project to build a new version of its pit stop analysis system.

2. (Acquisitions) Find sentences containing verbs {acquire, buy, take over} and return the subject and object of the verb if both are named entities of the type ORGANIZATION.

Example: Zensar acquires New York-based Thoughtdigital.

3. Find sentences containing verbs {join} and return its subject and object if they are named entities of type PERSON and ORGANIZATION respectively.

Example: Mr. Shah has joined Tangent as Vice President Human Resources.

Note: subject and object usually considered as semantics roles.

2.3 The INX Tool

Tata Research Development and Design Centre (TRDDC), Pune, India, has developed an IE Tool called INX. INX is available on MS-Windows platform. We have already explained in Section 2.2 the salient technology incorporated into INX. Important features of INX are as follows.

- **Powerful semantic analysis**: INX provides a powerful semantic analysis facility. It not only understands the constituents of sentences but also resolves the relationship between the constituents. In addition, INX identifies named entities such as organization, person, location.

- **Gazette creation**: Gazettes are domain specific resources that INX requires to carry out extraction effectively. INX provides number of ways to create gazettes from the text.

- **User extensible pre-processing**: INX provides a comprehensive library of pre-processing routines that enable the text to be thoroughly cleaned before being further processed. Furthermore, INX allows the user to customize and extend the library to handle situations specific to the text and domain being processed.

- **Comprehensive training cycle support**: INX provides extensive support for managing training cycles. Apart from providing ways to create multiple training cycles it also provides a comprehensive set of reports that collect metrics within a training cycle and across training cycles to enable the extraction process to be effectively monitored.

- **Multi-document format handling**: Ability to read documents written in a variety of document formats such as Microsoft Office, HTML, and XML.

- **Interfaces to document databases**: INX interfaces with Microsoft SharePoint Document Server.

- **Multi-format output**: The information extracted from the text can be viewed as relational tables, XML or CSV files.

3. COMPETITOR INTELLIGENCE

We now discuss how we have used the INX Tool to analyze a set of financial news items and to automatically extract the details of a specific type of financial event. We first discuss events of the type mergers and acquisitions (M&A). We have already shown in
Section 1 an example of the output of INX which is stored in a template (or database table).

More details about the mergers and acquisitions events (e.g., date, locations, value etc.) can be acquired as well. Mention of this financial event in a sentence is recognized by verbs such as merged, acquired, bagged etc. These characteristic verbs are stored in a gazette in Inx. More verbs can be added to this gazette if necessary. Once a sentence is identified (e.g., by the occurrence of one of these verbs), then Inx identifies whether any organizations are mentioned in the sentence and the role played by them vis-à-vis the verb. For example, the sentence “Zensar Technologies has acquired ThoughtDigital, a New York-based software firm that specializes in Oracle applications.” contains an organization “Zensor Technologies” which is the subject of the verb “acquire” and another organization “ThoughtDigital” which is the object of the same verb. Thus Inx infers that Zensar Technologies has acquired ThoughtDigital and adds an appropriate record to the output table.

Inx analyzes the given set of news stories using NER and semantic analyzer and identifies organizations and semantic roles played by these organizations for actions such as merged, acquired, bagged etc. Inx combines the information available by both the processes - that is, semantic roles for the phrase and its respective named entity tag – and generates text enriched with all the semantic information available. Inx uses simple regular expression using these relational (action-actor) and NE tags to extract the desired information. The sample regular expression, if written in English looks like:

Extract subject and object of the verbs {merge, acquire, …}, provided subject and object are Organizations.

Every news story is processed in sequence and information is extracted as required.

In a similar manner, we have programmed Inx to extract details of another type of financial event: deals and contracts from the given set of news items.

Fig. 6 shows two example news items reporting award of projects and contracts. INX obtains the following structured data by analyzing these news items.

<table>
<thead>
<tr>
<th>Company bagging the deal</th>
<th>Value</th>
<th>Company from</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satyam</td>
<td>Dollar 200 million</td>
<td>Applied Materials Inc</td>
</tr>
<tr>
<td>Cognizant</td>
<td>-</td>
<td>Simon and Schuster Inc</td>
</tr>
</tbody>
</table>

The accuracy for mergers and acquisitions as well as deals and contracts on various news items is above 90% - that is, Inx was able to extract more than 90% of the mergers and acquisition events mentioned in the given set of news items.

![Image](http://www.x-media-project.org)

**Figure 6. Two sample news items reporting award of contracts and projects.**

4. RELATED TOOLS AND CONCLUSIONS

We now describe a few IE tools that provide competitor analysis functionality: ClearForest (http://www.clearforest.com) provides a built-in module for finding mergers and acquisitions, but not other kinds of financial events. TEMIS (http://www.temis-group.com) provides a built in module (called skill cartridge – comprising domain specific extraction rules and dictionaries) for competitor analysis. Unlike ClearForest, it also extracts joint ventures, stock purchases, innovations, sales information, stock information etc. Biovista (http://www.biovista.com) a supplier of corporate intelligence services in biotechnology domain, is working on a product for competitor analysis. The x-media consortium in Europe (http://www.x-media-project.org) will provide competitor analysis as part of their multi-media knowledge management solution.

In this paper, we have presented the INX information extraction tool and discussed its multi-pronged approach to extraction. To summarize, INX creates an enriched representation of the original text, by adding results of syntactic analysis, semantic analysis (identification of semantic roles) and named entities. Usual regular expressions are then used to search this enriched text representation and to extract the desired information from it. The extracted structured information is added to relational tables (or XML files). We then discussed how we have used the INX Tool to extract competitor information (i.e., details of the financial activities and events related to competitor organizations) by analyzing news items. We gave examples of events like mergers and acquisitions and award of contracts.
INX has successfully demonstrated that it can extract details of financial events and activities by analyzing news items. We are now looking at extracting details of many more types of financial events. Improving the accuracy of extraction and making the extraction patterns more robust to different news sources are also issues we are looking at. We also wish to look at the problem of applying statistical and data mining techniques to this extracted data and discover significant knowledge about the technical and business strategies of competitor organizations. For example, perhaps we may be able to discover that company A is currently focusing on getting financial projects in the Middle-East, rather than its previous strategy of manufacturing projects in Europe. We may find interesting patterns in mergers & acquisitions also.

Acknowledgements. We wish to thank our colleagues in TCS Pune and Mumbai for useful feedback and discussions.

REFERENCES