Acquiring Competitive Intelligence from Social Media

Lipika Dey, Sk. Mirajul Haque, Arpit Khurdiya, Gautam Shroff
TCS Innovation Labs, Delhi
{lipika.dey, skm.haque,arpit.khurdiya, gautam.shroff}@tcs.com

ABSTRACT
Competitive intelligence is the art of defining, gathering and analyzing intelligence about competitor's products, promotions, sales etc. from external sources. The Web comes across as an important source for gathering competitive intelligence. News, blogs, as well as social media not only provide competitors information but also provide direct comparison of customer behaviors with respect to different verticals among competing organizations. This paper discusses methodologies to obtain competitive intelligence from different types of web resources including social media using a wide array of text mining techniques. It provides some results from case-studies to show how the gathered information can be integrated with structured data and used to explain business facts and thereby adopted for future decision making.

Categories and Subject Descriptors
H.3.1. [Information Systems]: Information Storage and Retrieval. Content Analysis and Indexing. Linguistic Processing

General Terms
Content Analysis, Design, Experimentation.

Keywords
Text Mining, Competitive Intelligence, Business Intelligence from Social Media

1. INTRODUCTION
Competitive intelligence is defined as a combination of defining, gathering and analyzing intelligence about products, customers, competitors and any aspect of the environment needed to support executives and managers in making strategic decisions for an organization. As opposed to industrial espionage, competitive intelligence is viewed as a legal business practice with focus on the external business environment. It involves defining a set of processes to gather information, converting it into intelligence and then utilizing this in business decision making. Business users also emphasize on the usability and actionability of the information gathered.

Competitive intelligence is aimed at assessing risks and opportunities in a competitive environment before they become obvious. Experts also call this process the early signal analysis. Gathering competitive analysis is a highly specialized activity and difficult to automate. However, given the changing landscape of business environments and the wide-spread availability of a large array of information on the web, it has become a necessity to design tools that can aid analysts in the competitive intelligence gathering and analysis process. Utilization of the knowledge and coming up with effective business strategies is not in the ambit of these systems. Listening platforms typically engage in listening to the social media to gather consumer pulse. Competitive intelligence is not restricted to gathering knowledge about the consumer only. It studies expert opinions, technology advancements, economic policies, social changes and many other parameters essential for excelling in business.

Traditionally, analysts relied on a whole lot of facts and figures, often provided by third-party reports and survey results to gather competitive intelligence. A large amount of information required for competitive intelligence is unstructured in nature and therefore not immediately machine-interpretable. The availability of many of these resources on web ensures that the volume of such information is humongous. News generated from multiple sources also contributes to this collection. Discussions on different forums and blogs can provide crucial competitive intelligence when analyzed in proper perspective. Social media content can also contribute to competitive knowledge in a big way. Appropriate knowledge management techniques within an organization can ensure that all relevant internal information is available easily to analysts to impact decision-making in a positive way.

Though a large volume of information is received in a digitized format, it requires sufficient human effort to read, extract, organize and assimilate relevant knowledge from these sources. The complexity of the task arises from the heterogeneity of sources as also the non-standard presentation schema. Given the volume and rate of information accumulation, extraction of relevant information from the collection and its assimilation requires substantial human effort. Variation in relevance and reliability of web-data also add to the uncertainty of the process. The noise in web content is a major issue that has to be dealt with before using the content. Noise in this case has many connotations. While one type of noise arises from the use of non-standard language and vocabulary, a different kind of noise appears due to the fact that the context of a piece of content cannot be easily ascertained, though it is important to interpret the content in the context of its occurrence. Context assessment is a generic problem faced during machine interpretation of unstructured data. The problem increases manifold when the task is to interpret unstructured content from a web-page, since the context changes dynamically.

In this paper, we first try to categorize the different aspects of competitive intelligence that can be gathered from web-

J-MOCR-AND '11, Beijing, China.

documents. Thereafter, we attempt to identify and characterize different types of web-resources that can provide such knowledge. Finally, we present how text-mining can help in acquiring this knowledge. While it is not our aim to state that the process of acquiring and extracting competitive intelligence can be entirely automated, we show how text-mining techniques can be employed to design a set of tools to aid the analysis tasks. We present methodologies for extracting and categorizing different types of information from text data and converting them to usable chunks of information. We further present some case-studies through which we show how the extracted information can help in decision making for business.

Starting with a detailed discussion on different types of resources and the competitive intelligence that can be gathered from these, we thereafter focus on a subset of these elements which can be gathered from social-media data and news.

2. REVIEW OF RELATED WORK

Marketing research, which was a pre-cursor to competitive intelligence gathering, is a tactical, methods-driven field that was almost solely dedicated to primary research that analyzed customer data gathered through surveys or focus groups. It aimed at unearthing beliefs and perceptions about companies using statistical techniques. Competitive Intelligence typically draws on both primary and secondary sources from a vast range of stakeholders inclusive of suppliers, retailers, media, and competitors and so on [1]. Competitive intelligence focuses on analyzing the present problems as well as project into the future.

A set of basic requirements for competitive intelligence was laid down in [2], which also highlighted the differences between CI and market research. CI has a specific perspective of calculating external risks and opportunities to the firm's overall performance, and hence different from other information activities. In his book [3] Rappaport presents a number of case studies that show how social media conversations can be turned into business advantage.

While CI still remains largely perspective-driven, the volume of available information has grown manifolds due to the Web. De Olieviera et al. [4] presents how text mining techniques can be used to analyze an enterprise's external environment searching for competitors, related products and services, marketing strategies and customers' opinions. They reported a case-study that uses domain knowledge to extract concepts from the texts and then search for pre-defined patterns. Patterns were category-specific and primarily defined by required presence or absence of concepts. In [5], Zanasi has compiled several articles that describe how text mining techniques can be applied to extracting consumer intelligence, along with case studies from various industries. Patent mining using text mining techniques has been around for quite some time now. [6] describes a series of text mining techniques used by patent analysts. These techniques include text segmentation, summary extraction, feature selection, term association, cluster generation, topic identification, and information mapping. [16] reports use of statistical inference mechanisms for mining business entities and business relations from text. Analyzing consumer sentiments and opinions is yet another mature application area of text mining. Though sentiment analysis is fairly subjective [7] and accuracy levels are yet to be very high, the role of text mining in improving the recall value while analyzing large volumes of text can be hardly over-emphasized. [8-10] describe several different approaches to mining opinions and sentiments from noisy text data.

Traditionally, enterprise Business Intelligence still works with structured data only. [11] and [17] discuss how unstructured data analysis can be integrated into the Business Intelligence framework.

3. CHARACTERIZING COMPETETIVE INTELLIGENCE

Organizations use competitive intelligence to compare themselves to other organizations ("competitive benchmarking"), to identify risks and opportunities in their markets, and to pressure-test their plans against market response (war gaming), which enable them to make informed decisions. Most firms today have substantial online presence and it is possible to gather knowledge about what the competitors are doing and how the industry is changing. The information gathered allows organizations to realize their strengths and weaknesses. Information gathered in the process can be interpreted in different ways by different executives depending on their perspective for analysis.

Competitive intelligence can be broadly classified into two categories depending on whether it is used for long-term planning or short-term planning. Strategic Intelligence (SI) focuses on long term issues that analyze a company's competitiveness over a specified period in future. The actual time-line depends on the type of the industry. The main focus of analysts here is to forecast where the organization should be positioned in x Years and to identify strategies to convert this into a reality. This type of analysis primarily involves identifying weaknesses and early warning signals within the organization. Tactical Intelligence focuses on providing information that can influence short-term decisions. Most often, this is related to analysis of current market share and the competition landscape. This kind of intelligence directly affects the sales process of an organization.

Tactical intelligence can be further categorized as follows:

(i) Market information – provides information about popularity of competitors in terms of their products or brands as a whole, which products are moving in the market, market share of competitors. Regional or geographical biases of competitors also fall in this category. Consumer sentiments related to the organization and its competitors also belong to this category.

(ii) Price information – provides knowledge about prices of competitor products.

(iii) Promotion – Provides information about promotion strategies and kind of promotional activities that are adopted by competitors.

(iv) Other issues – Organizational information about competitors like their work force structure, internal shift in focus or vision, success or failure of their trials, new product launches, technology investments etc. all contribute towards building a profile of competitors that can be useful to organizations.

Competitive intelligence when gathered and analyzed in a timely manner can help organizations to substantially reduce their reaction times. Responses by an organization may be in the form of price adjustments, changing marketing strategies, revising production plan etc. Several major airlines constantly track competitor fares and readjust their prices in very short notice.

Analysis reveals that different types of Internet resources may provide different types of competitive knowledge. Internet sources like Hoover's Online service can provide information about company profiles and financial information integrated from millions of public and private companies for diverse set of industries. Social media can provide information about brand popularity, consumer sentiments and competitor promotions. News, discussion forums and blogs can provide information about other issues like technology investments, product launch, or vision announcements by competitors. Tracking online trading sites can provide information about competitor promotions and brand sentiment. Table 1 captures the different kinds of web resources along with the types of competitive intelligence they can provide.

Table 1: Competitive Intelligence resources on the web

Type of Competitive Intelligence	Web resources
People events	News, company web-sites
Competitor strategies – Technology investment etc.	News, Discussion Forum, Blogs, patent search sites
Consumer sentiments	Review sites, social networking sites
Promotional events and pricing	Twitter, Facebook
Related real-world events	News, Twitter, Facebook

4. GATHERING COMPETETIVE INTELLIGENCE FROM THE WEB

The web provides an avalanche of information, which needs to be gathered and processed before analysts can use the information effectively. In this section we propose the design of a dedicated, virtual business intelligence system that delivers intelligence-digests to decision makers in a timely manner and in a format that enables them to make decisions pro-actively. The design is generic and can be extended to any industry with appropriate customizations.

Figure 1 presents the design of the proposed system. It has several modules responsible for gathering, cleaning, processing, harmonizing and consolidating intelligence information from a wide array of sources.

The content acquisition and assimilation module is responsible for gathering and extracting content from an array of sources.

i. A set of site-specific focused crawlers are deployed to extract content from a set of pre-defined sources including popular social networking sites. The list of such sites is supplied as input to the system. Social media sites like Twitter offer APIs like garden-hose or fire-hose to collect relevant tweets.

ii. In order to ensure that no relevant information is missed even if it is posted on a new or insignificant site, the system also employs Google Alerts[1] to improve recall. Google Alerts takes in as input a set of key-words or phrases and in return provides the system a list of links to the sources that contain relevant content. Key-words and phrases include competitor names, product or service names etc. The system then deploys the open-source content crawler called Nutch2 to extract content from these web-pages. Repeated as a periodic exercise, this method ensures that even insignificant events reported on the Web is definitely obtained by the present system as soon as they start receiving customer hits due to Google's relevance computation method.

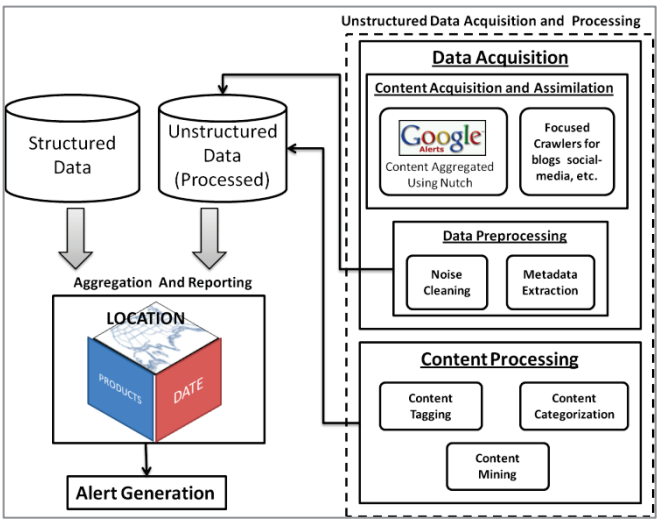

Figure 1: System for gathering online Competitive Intelligence

The data pre-processing module is responsible for cleaning and extracting relevant content from different sources. Noise removal involves two steps:

(i) Obtaining clean content from web-sites – Since each web-page has its own characteristics and contains unwanted material like advertisements, or links to other irrelevant web-pages etc, the system has to learn page-specific rules to extract relevant content from web-pages. These rules are learnt over time by analyzing web-sites and identifying the static and the dynamic structural elements. The static structural elements usually contain links. menus etc. which are irrelevant for the purpose. The dynamic elements are further analyzed to differentiate between advertisements and content. Advertisements usually contain images, videos, links etc. and are rejected. The text content that changes dynamically within a page is finally selected for further processing. HTML tags like title, author, heading, by-line etc. are used to generate associated meta-data. Place association is not always straight-forward. For news sources, it is usually at the head of the content followed by a standard de-limiter or date.

(ii) Cleaning consumer generated text – While news text is fairly clean, content on social-networking sites, blogs and discussion forum are fraught with noise. The system employs several techniques, some of which were presented [10] to clean the content. These include context-dependent spelling correction, sentence demarcation, removing unnecessary capitalization and special characters etc.

[1] http://www.google.com/alerts

[2] http://nutch.apache.org/

The cleaned content is then stored in a central data-store along with all available meta-data information like author, publish-date, place of reporting, type of source, thread of conversation etc. All meta-data fields may not be applicable to all types of content.

The content processing module has the most important responsibility of identifying and tagging the relevant content from the vast collection. It consists of several sub-modules. The sub-modules implement an array of NLP and text-mining methods to accomplish its tasks. Content assimilation and categorization are largely driven by business requirements. A well-structured business knowledge-base, domain ontology and different types of dictionaries for linguistic processing are used by this module. The nature of information required may be unique to an organization, though there are sufficient commonalities within a domain. One of the key features of the proposed system is the ease with which it can be customized to the requirements of an organization.

All classified content is harmonized and consolidated to generate reports in pre-defined templates. The process of consolidation is knowledge-driven. Consolidated reports are quantifications of extracted information and can be treated as intelligence-digests, which provide analysts an access point into the underlying data for further exploration through drill-down facilities. Report templates are defined using company-specific context to deliver the information. Consolidation proceeds along three major dimensions. These are (i) time (ii) location and (iii) product and services. All information pertaining to a single product or service are aggregated together, sorted by time and segregated by place while presenting to the end-user. Organizational knowledge base contains a detailed hierarchy about products and services and is used for the purpose. It may be noted that acquired information components may have variable levels of granularity attached to them. The information presentation layer takes the granularity into account while presenting the reports to the end-user. Reports are actionable and aimed at enabling decision-makers to act in an informed way.

The alerting module assesses different kinds of processed information and generates alerts on pre-defined triggers or "hot" events that have been identified as preambles to scenarios that require immediate responses. One example of a trigger event is the news-reporting of a product launch by a competitor, along with details about the product. This may be a possible warning for lowering of sales or market share in subsequent quarters. The public response to the news gathered from different sources can be used to roughly evaluate the possible impact. Alerts are also accompanied by reports.

5. INFORMATION EXTRACTION FOR COMPETITIVE INTELLIGENCE

The first step in content processing involves identification and tagging of relevant content. The present system employs concept-based tagging. Concepts are defined in terms of words, phrases, entities and their combinations along with syntactical and grammatical restrictions imposed on the combination. Organizational information about its products, services and strategies can be very effectively stored in domain ontology. [12-13] describes methodologies for creating domain ontology from unstructured text. We have employed similar methodologies for creating domain ontology from web-site content, with humans in the loop. The ontology also defined mapping between different domain entities and generic terms for those as extracted from a thesaurus. This is done through the use of WordNet3. The system is also equipped with a rule-language to encode business rules.

While domain dictionaries provide an initial set of concepts, concepts relevant for a domain can also be learnt using machine-learning driven evolutionary platform, and eventually build a collection of concepts and patterns over time. A frequently occurring concept or combinations of concepts within a sub-set of documents can be used to label the sub-set, provided the pattern can be aligned to a competitive intelligence component. For example, a collection of documents can be labeled as related to "Merger and Acquisition". If a subset of documents is primarily about a specific competitor, it can be assigned the competitor label also. Finally, document collections can be tagged with specific competitive intelligence tags like "People event" or "Product Launch event" etc.

The NLP module performs entity extraction, phrase extraction (NP, PP) and grammatical relations extraction. Relations can be unary, binary or ternary. Unary relations involving adjective or adverb modifiers and entities represent entity or object roles. Binary relations involving subject-verb, or verb-object and ternary relations involving subject-verb-object represent different types of events. For relation extraction we have integrated our system with ReVerb – Open Information Extraction Software4, developed by University of Washington and available as open source.

The system is equipped with a range of temporal and spatial relations that can be used to constrain the relative times of occurrence and relative positions of occurrence among information components or contained entities and events. Generic spatial relations like *preceded_by, followed_by, within_distance_x_of_each_other* are positional in nature and operate on textual elements like entities or relations. Relations like *before, after, overlapping, starts_during* etc. work on time dimension and are constrained to work on relevant meta-data only. Each text content is associated to its publish date. Though the time of an event need not be same as the time of publishing, since time values extracted from within the text documents are not very accurate, we stick to the publish-date. All date factors including month, time, week etc. are possible to be defined and used. Organizations across the world have varying calendars. So long as the calendar is defined for the system, associations at week level or quarter level are not difficult. For example, for organizations that start their year in February, the first week of March is the 5th week. Temporal associations are extremely important in ascertaining the importance of events and their effects to other organizational activities.

Relational information components mined from text are further analyzed to be assigned competitive intelligence tags. Tag assignment is based on dependency analysis followed by analysis of associated subject, object and nature of verb. Since a verb itself may have synonyms, we have employed VerbNet5 to classify verbs. The mechanism of assigning different types of tags is further explained with examples for each category.

(i) People events – Events related to key market players like CEO's or CFO's of companies are termed as people

[3] http://wordnet.princeton.edu/

[4] http://reverb.cs.washington.edu/

[5] http://verbs.colorado.edu/semlink/semlink1.1/vn-fn/

events. Market players are identified as named-entities using Named Entity Recognition (NER) techniques. There are several NER tools. We use the Stanford NER[6]. Titles, roles and honorifics such as "Mr.", "CEO", "Vice President" etc. are also tagged. The system further deploys dependency based analysis to identify relevant information. Unary relations comprising of the honorific and a named entity connected by the MOD or the modifier relation are extracted. Binary relational components with persons as subject and reporting verbs like *said*, *announced*, *declared* etc. are identified as people events. Ternary relational patterns comprised of relational information components involving persons as subject and organization names in object and typical verbs like "*serves*", "*holds*", etc. are also categorized as people events.

(ii) Competitors' strategies – Market news, patents, etc. provide insights about investment strategies by competitors. Information about new product launches, mergers and acquisitions, new investments in technologies, opening of a new store, engaging a new vendor etc. come under this category. Each of these activities can be associated to specific sets of action verb classes. Relation-based patterns are useful for identifying the events. For example, effect verbs like *rise*, *fall*, *advanced*, *went up* along with sales as subject and numerical components in object are indicative of such reports. Objects may also contain money values. Similarly concepts around *invested*, *appointed*, *enlisted* etc. are aggregated into this category.

(iii) Consumer sentiments and opinions – Blogs, discussion forums and social networks abound in consumer sentiments. All large companies have Twitter and Facebook presence to reach out to a large audience. Mining these discussions provides valuable insights. We apply opinion mining techniques presented in [10] to gather consumer opinions of different issues. Reports on comparative sentiments about competitors are thereafter generated.

(iv) Promotion events of competing organizations – Promotional events can be very effectively caught from the web. While retailers and brands vie with each other for consumer attention, innovative promotion schemes like Groupon, Foursquare etc. use social networking to ensure consumer participation. In the next section we have presented examples of pattern-based detection of promotion events. It is possible for an organization to gain real-time insights into competitor promotions and adjust its short-term as well as long-term strategies. Typical verbs like "promotes" or "pushing" are observed in promotional material. Concepts indicating promotions include *sales*, *discount*, *offer*, *free gift*, *% off* etc. Promotions can be further categorized into "Celebrity-led promotion" or "product promotion", "promotion of a cause" etc. We have mostly worked with product promotion.

(v) Real world events – Real-world events not related to market intelligence directly can also affect business in many different ways. While catastrophes cannot be foreseen, the effect of catastrophes can be reduced by having real-time peek into evolving situations even at remote places of the world due to Twitter. It is proved that Twitter report on such events precede the best of news sources by almost 30 minutes. Twitter is particularly an important source of information for events that have no global significance, but are very local in nature. These kinds of events, like a flash flood in a small locality or a factory fire may not be even reported in News but may affect the operations of either the organization or its competitors in a significant way. [11] presents how such tweets can be identified and analyzed to take pre-emptive steps and minimize losses for a supply-chain scenario. Patterns for this set cannot be pre-defined. Frequently observed new patterns are regularly added to the collection.

6. ASSIGNING COMPETITIVE INTELLIGENCE LABELS

Documents are assigned a Competitive Intelligence (CI) label based on their content inclusive of the entities, relations, concepts and opinions. Since articles are variable in nature the system follows different approaches for labeling them. For news articles and blog or discussion-forum text that are expansive in nature and usually grammatically more consistent, NLP-based techniques are employed to identify entities and relations. CI labels are based on the class of relations obtained. Social-network texts on the other hand are usually short and noisy in nature. Word and concept-based tagging are employed for assigning labels to these types of documents. Besides, news articles are mined for people-events and competitor strategy events, while social-network content is mined for opinions and promotion events.

While social-network content can be naturally grouped based on the context in which they appear (for example in reply to a specific comment), clustering helps in grouping related news documents and reducing the dimensions of analysis. The present system employs topic-based clustering on news articles to group them together. Topics are extracted using Latent Dirichlet Allocation (LDA) [14]. Each topic represents a group of news articles each of which maximizes that topic. Named entities and relations are extracted from all news articles. Most significant Named Entities and relations are then mined from each cluster using Maximum Entropy based computation as specified in [15]. The main verbs in the significant relations are further used to assign a CI label to the cluster and thereby all articles associated to it.

Social-network contents on the other hand are short and noisy. So labeling of these articles is rule-based. The system uses prior knowledge about enterprises, their products and other concepts like discount or promotions etc. to tag articles. Opinion mining is also another activity performed on this set. Each comment is thereafter associated with a positive and a negative opinion score along with the product concepts that they represent.

[6] http://nlp.stanford.edu/ner/index.shtml

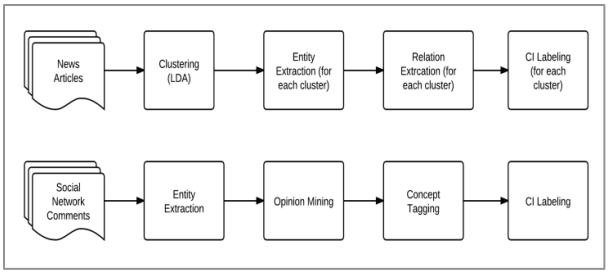

Figure 2: Approach to CI labeling for different types of documents

Figure 2 presents overview of the approach while Algorithm 6.1 presented below describes the details of the tagging process.

Algorithm 6.1

Let D be the collection of all unstructured documents at time t.

$$D = \prod_{d \in N,B,C} d$$

where N denotes the collection of news articles, B denotes the collection of blog and discussion forum text articles, and C denotes collection of short comments collected from social networking sites like Twitter, Facebook etc.

News Processing

- Collect all news articles stamped with the same time-stamp into a single group.
- Apply LDA based clustering on each group.
- Each cluster c obtained for a single d is assigned the same date-stamp as the articles contained in it.
- For each cluster c obtained,
    - Extract entities.
    - Extract relations using ReVerb
    - Label each relation as People-event or competitor strategy based on relation type or verb class.
    - Find the most frequent entities and relation labels in cluster c.
    - Tag cluster c by frequently occurring relation labels and entities that exceed specified thresholds.

Blog content Processing

For each blog site, identify the site as review site or discussion site.
For discussion sites, process articles like news articles.
For review sites, process articles like comments as described next.

Social-Network content processing

- For each comment collected from a social-networking site the comment has the following meta-data associated to it
    - -author
    - -time of publishing
    - -comment heading – this is same as the comment if it is a new comment. However, if the present comment is a reply, replica or an extension of an earlier comment, then the comment heading stores the original comment.
    - -comment content – the actual text of the comment
    - -external links – In case a comment contains an URL to an external content, the URL is presently stored in this field. The present system does not process this content any further.
- Perform entity extraction from comment content.
- Perform opinion mining on comment content.
- Perform concept-based tagging on comment content. Assign product or service category of promotion through ontology look-up for entities and concepts.
- Assign CI label
    - Assign label - "promotion event" if comment contains concepts related to promotion
    - Assign label OPINION – if comment contains opinion

Aggregate all articles into database along with meta-data like time and associated tags like entity labels, Competitive Intelligence event tag and opinions.

End Algorithm.

We now present some results in Table 2 that show sample relational components extracted from news articles and the CI labels assigned to these components by the system based on these relational components. These news articles were collected during the period April – June 2011 for the retail domain, by setting up alerts for three leading retail stores of USA. A News article may contain relations that contain subjects other than those used for alerting. Social networking sites for these three retailers were also crawled for consumer-generated data. The data typically contained user comments in response to posts by the store owners. Store-owner posts either announced upcoming or ongoing sales or were targeted at generating user reactions to specific issues. Garden hoses were set up to collect tweets containing mention of these stores. Table 3 shows results of analyzing these tweets.

Table 2: Relations identified from news and corresponding CI event Tagging

Cluster#	Dominant Pattern	CI Label
11	<Retailer, Verb-phrase with key verb open, object of class Organization>	Opening of new store
Apr, 25, 2011: Macy's to open 3 new Bloomingdale 's outlets Apr, 27, 2011: The Dillard 's property opened doors to Westgate 's redevelopment May, 10, 2011: Sephora has opened in Malaysia. May, 23, 2011: Cincinnati-based Macy 's is opening two new stores		
Cluster#	Dominant Pattern	CI Label
21	<Named Entity, Verb-phrase with key verb launch, object class PERFUME>	Product Launch - (Perfume)
May, 07, 2011: Kim Kardashian celebrated the launch of her second fragrance May, 08, 2011: Jennifer Aniston has launched her new signature fragrance. May, 25, 2011: Laura launched her own home fragrance collection		
Cluster#	Dominant Pattern	CI Label
0	<Organization, Key verb in verb phrase = acquire, Organization>	Acquisition of company
Apr, 05, 2011: Teva Pharmaceutical Industries Ltd. plans to acquire Cephalon , Inc. Apr, 07, 2011: Britain's Sports Direct has acquired iconic U.S. boxing glove manufacturer May, 23, 2011: BzzAgent has been acquired by dunnhumby Ltd.		
Cluster#	Dominant Pattern	CI Label

6	<Organization, Key verb in verb phrase = promote, Product>	Promotion of products
Apr, 29, 2011: P&G heavily promoted Pampers Swaddlers. May, 29, 2011: J.C. Penny online is promoting furniture and mattresses.		
Cluster#	Dominant Pattern	CI Label
20	<Named Entity = Person, "Key verb in verb phrase = serve, concept = POSITION IN ORGANIZATION>	People event
Apr, 06, 2011: Mr. Martin also served as CSI's Interim Chief Financial Officer Apr, 18, 2011: Vitale currently serves as Chairman of the Academy Apr, 21, 2011: Richie served most recently as Senior Vice President and Chief Marketing Officer		
Cluster#	Dominant Pattern	CI Label
3	<Subject contains Sales, key verb in verb phrase = class(affect), object contains numerical followed by %>	Rise or fall of sales
Apr, 07, 2011: March same-store sales fell 6.5% May, 06, 2011: Consumer domestic sales edged up 0.9%. May, 09, 2011: Online sales were up 38.3%. May, 21, 2011: Sales of motor vehicles and parts fell 2.4 percent May, 25, 2011: First quarter net sales fell 9.8%		

Table 3 presents a sub-set of tweets that were classified as promotional offers and categorized item-wise from a total of 181000 collected for seven days. The re-tweets and references to the same tweet have not been shown here. The data has been re-produced as it is without much editing other than cleaning.

Table 3: Tweets tagged with CI label "Promotion Events" and product category based on product concepts

Date	Product Category	Tweet
08-06	Perfume	Palazzo Italia Online Boutique Shop Sale Cheap Dolce&gabbana Light Blue Women Perfume http:// dlvr. it/ VVJNQ.
09-06	Perfume	POLO BIG PONY 1 cologne by Ralph Lauren for Men : Discount Discount Designer Perfumes, Colognes and Fragra.
07-06	Perfume	Hey ladies, check this Special discount for Gucci ?Guilty? perfume only Matahariteptstore and save 20%, until.
09-06	Perfume	2011 Xmas Sales- Polo Sport by Ralph Lauren for Men, Eau De Toilette Natural Spray, 1.3 Ounce: Perfume Polo Spor.
10-06	Perfume	EAU DE Discounts D&G Womens Fragrance - D & G 3 L'I.
11-06	Perfume	GUCCI is Up for Sale, save $35.99 and buy it now for $73.
11-06	Perfume	Shop here http://bit.ly/mKalPL Perfume Scent Fragrance Australia.
11-06	Perfume	Tracy Morgan said storestoavoid is Wal-Mart / Macy's for Perfume & Cologne. Instead Visit Fragrancevelly for Free Shipping in USA Sale.
11-06	Perfume	Fetus' perfume is for sale at Sephora http:// bit. ly/ lfJdAM.
13-06	Perfume	Ralph Lauren Polo Blue Cologne and Perfume on Sale with Free Shipping in USA
		http://t.co/NkmEylw.
14-06	Perfume	Classic Brown Box Gucci Eau De Parfum for Women 2.5 Oz EDP perfume NIB Sealed http://t.co/qgRzJGG Free Shipping in USA SALE.
14-06	Perfume	Gucci Womens Clothing On Sale: $27.50 Gucci By Gucci Perfumed Shower Gel 200ml/6.7oz http:// bit. ly/ iMH3Td.
06-06	Dress	Macy's sale: Up to 88% off women's plus size dresses: Macy's takes up to 88% off a selection of women's plus si.
06-06	Dress	Macy's Coupon: 25-40% off Plus-size Dresses http:// dealspl. us/ macys- coupons/ 272770.
07-06	Dress	RT DealsTipper: Macy's great Summer dress sale with 25 - 40% OFF.
10-06	Dress	Gucci dresses on sale http:// dlvr. it/ Vk2fD.
12-06-2011	Dress	Diesel jeans for sale - bebe b dresses /b under $100 - gucci woman shoes http://bit.ly/kvGEDK clothing.
12-06	Dress	Cheap karen millen dresses:karen millen sale Sample Sale Scoop Brian Reyes, Whitney Eve, Gucci Vintage, More http:// bit. ly/ lQwUaC.
13-06	Dress	Men's Ralph Lauren Dress Shirts: For sale are several Ralph Lauren Men's dress shirts.
13-06-2011	Dress	Macy's - $25 off Bridesmaid Dresses coupons http:// buxr. com/ d/ 85396.
13-06	Dress	Im j rose yea Dillard's has a dress sale today.
14-06	Dress	Sale ralph lauren women dress http:// goo. gl/ fb/ zINt8.

6.1 Learning Competitive Intelligence Labels

The present system is mostly rule-based and therefore not very scalable. However, we propose to integrate machine-learning mechanisms to convert this to an evolving system. Starting with a set of documents that are marked by the rule-based system, the system can mine for frequent patterns within each labeled collection. The patterns can be ratified by human experts and editable. All available metadata, along with extracted linguistic elements like words, entities, phrases or relations are available for defining the patterns. Human experts can add or delete additional constraints to a pattern. Patterns can be viewed as rules that combine positive and negative occurrences of expressions indicating the presence or the absence of information components and their combinations in a document. Finalized patterns can be incorporated into the rule-base to enhance it. The patterns are thereafter converted into internal queries to label the documents automatically, based on the patterns visible in them. A complete rule language has been specified to accommodate complex queries involving relational, spatial and temporal operators. The system uses the queries to label future data. Patterns can also be used as features for training a classifier to automatically classify new documents as they come.

7. CASE STUDIES

Competitive intelligence reports can be generated and compiled according to various conditions. Grouping of reports is often dependent on the perspective of analysis. For example, a report which states the launching of a new electronic product in the market by a rival company, may lead a section of analysts to study consequent promotion events and their effects on market share.

Another section of analysts may be interested in studying the technological aspects. We have conducted some preliminary studies for a few organizations about how social-media data and news articles can be used to derive competitive intelligence and integrated with internal data. One of the aims was to study the correlation of rival brand promotion events on sales data. The other aim was to study the feasibility of using Social Networking sites as sources for consumer sentiments. We present some results without naming the organizations or products for reasons of confidentiality. It was observed that event location information is not reliably available from Social Network Data. While integrating information this dimension was ignored.

Observations from the study are summarized below:

(a). Social media data can be an indicator of brand popularity – Facebook data for the stores mentioned earlier was analyzed to compute brand popularities for different well-known brands for perfumes. Figure 3 shows week-wise variation in brand popularity, computed as a function of positive sentiments posted for a brand and volumes of comments by consumers. It was observed that popularity of a brand was directly correlated to promotional activities for the brand, as high-lighted in Figure 3. Figure 3 shows promotion announcements for a cosmetics brand called Clinique as white call-outs. Interestingly, popularity of this brand was directly correlated to its promotion which went through weeks 2 to 8. The brand as such did not figure when consumers were directly asked to vote for their favorite perfume brands subsequently in week 9.

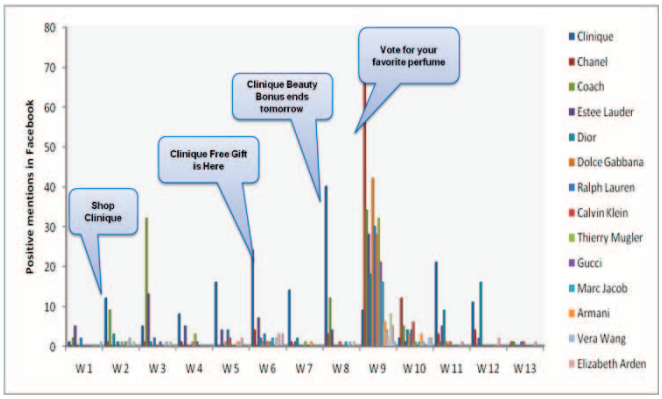

Figure 3: Brand popularity in Facebook superimposed with promotion events shows direct correlation between the two

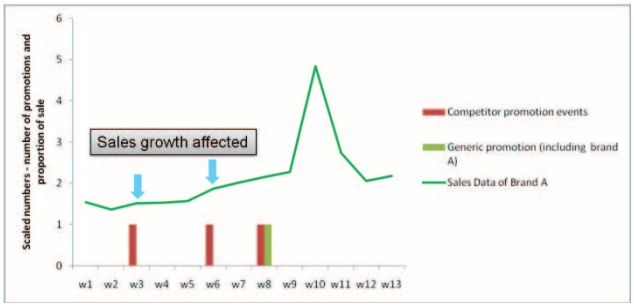

Figure 4: Correlating sales data with promotion events shows sales is affected whenever there are competitor promotions

(b). Correlation between social-media events and sales data of a particular brand – Figure 4 shows super-imposition of real sales data of a single brand with social-media events addressing all brands. It was observed that promotion of a brand led to either reduction of sales or at least arrested the growth in sales of rival brands. However, the effect was not observed uniformly across all stores and all geography. The impact was more in some regions than others.

(c). Cumulative impact of promotion events and negative news about competing brands on a particular brand – A preliminary study revealed that market share of a brand went up when its rival organization had announced a price rise. The price-rise event was caught from the news.

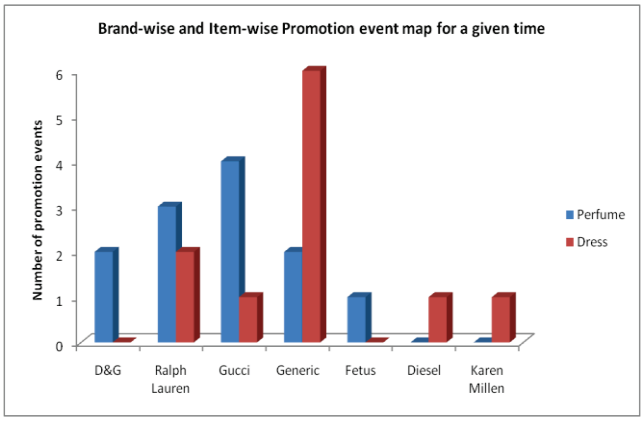

Figure 5: Promotion event mapping from Tweets

(d). Effects of different types of promotion activities on sales – More detailed study on this is yet to be done. This needs large volumes of past data to be correlated with current data. Companies can use the knowledge to design effective marketing strategies in future. However, as a first step towards this analysis, the data collected by the present system can be used to build a promotion event map for all competing products. Figure 5 presents a sample promotional event map generated from data shown in Table 3. This graph provides idea about relative volumes of promotions run by different competitors.

(e). Effects of different categories of news items on the performance of an organization. This is yet another area where large volumes of present and past data are needed to come to some conclusions. Machine learning strategies can segregate the relevant from the irrelevant information for large collections. Data collection for the process is on.

Promotion event maps and product launch information can be used by organizations for deciding their own short-term product-promotion strategies. We are currently building a predictive framework where past observations are used to generate alarms in case a sales dip is predicted due to upcoming promotions. Similarly, social media responses to sales announcements can be used as predictors for large sales volumes also. This can effectively reduce negative comments related to non-availability of discount gifts etc. eliminated.

8. CONCLUSIONS

In this paper, we have presented some preliminary results on our text-mining based system for gathering competitive intelligence from the web. The web abounds in information about competing products and companies. While the information is very valuable for analysts, getting access to the data in a timely fashion is a problem. Given the volumes of data that is there on the web, it is impossible for any human to go through it and make sense without aids for doing so.

Our target is to build a competitive intelligence gathering and processing system that can be customized to collect different types of data from the web. The data is processed and stored locally for further processing. Processing is knowledge-based and statistical. Preliminary results are promising. Correlation with structured data like sales data shows that it is possible to use this data and derive a lot of intelligence on how competition is affecting business.

This paper reports only some preliminary studies. As the results are combined with real-world business data, more avenues for analysis are likely to come up. The future focuses are on automating the reasoning process of linking external and internal data and generate intelligent alerts. Emphasis will also be given to assessment of data quality.

9. REFERENCES

[1] Ben Gilad and Jan Herring, "CI Certification – Do We Need It?", Competitive Intelligence Magazine, 2001, 4(2), 28-31.

[2] D. Blenkhorn and C.S. Fleisher, Competitive Intelligence and Global Business. Westport, CT: Praeger, 2005

[3] Stephen D. Rappaport, "Listen First – Turning Social Media Conversations into Business Advantage", John Wiley & Sons, Inc., 2011.

[4] José Palazzo M. de Oliveira et al., Applying Text Mining on Electronic Messages for Competitive Intelligence, 5th International Conference on Electronic Commerce and Web Technologies, EC-Web 2004, Zaragoza, Spain 2004.

[5] A. Zanasi, Text Mining and its Applications to Intelligence, CRM and Knowledge Management, Southampton : WIT Press, 2005.

[6] Yuen-Hsien Tseng, Chi-Jen Lin and Yu-I Lin, Text mining techniques for patent analysis, Journal of Information Process & Management, Vol. 43, 5, 2007.

[7] Bing Liu, Sentiment analysis and subjectivity, in Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010

[8] Bo Pang and Lillian Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, Vol. 2, 1-2, 2008.

[9] Nitin Jindal and Bing Liu, Identifying comparative sentences in text documents, SIGIR '06 - Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006.

[10] Lipika Dey and Sk. Mirajul Haque, Opinion Mining from noisy text data, International Journal of Document Analysis and Recognition, September 2009.

[11] Gautam Shroff, Puneet Agarwal and Lipika Dey, Enterprise Information Fusion for Real-time Business Intelligence, FUSION 2011, Chicago, July, 2011.

[12] M. Missikoff, P. Velardi, P. Fabriani, Text mining techniques to automatically enrich a domain ontology, Applied ntelligence 18 (3) (2003) 323–340.

[13] Missikoff, M., Velardi P., and Peterson, L. L. Reasoning about naming systems. ACM Trans. Program. Lang. Syst. 15, 5, 1993, 795-825.

[14] Blei et al., 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," The Journal of Machine Learning Research, vol.3, pp. 993–1022, 2003.

[15] Arpit Khurdiya, Lipika Dey, Nidhi Raj and Sk. M. Haque, Multi-perspective linking of news articles within a repository, to be presented at IJCAI 2011, Barcelona, Spain, July, 2011.

[16] Raymond Y.K. Lau and Wenping Zhang, Semi-supervised Statistical Inference for Business Entities Extraction and Business Relations Discovery, SIGIR 2011 workshop, July 28, Beijing, China, 2011.

[17] Henning Baars and Hans-George Kemper, Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework, Information Systems Management, 25: 132–148, 2008.