MINING CONSUMER TRENDS FROM ONLINE REVIEWS:
AN APPROACH FOR MARKET RESEARCH

by

Olga Tsubiks

Submitted in partial fulfilment of the requirements
for the degree of Master of Electronic Commerce

at

Dalhousie University
Halifax, Nova Scotia
August 2012

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "MINING CONSUMER TRENDS FROM ONLINE REVIEWS: AN APPROACH FOR MARKET RESEARCH" by Olga Tsubiks in partial fulfilment of the requirements for the degree of Master of Electronic Commerce.

Dated: August 10, 2012

Supervisor: _____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: August 10, 2012

AUTHOR: Olga Tsubiks

TITLE: MINING CONSUMER TRENDS FROM ONLINE REVIEWS:
AN APPROACH FOR MARKET RESEARCH

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: MEC CONVOCATION: October YEAR: 2012

_____
Signature of Author

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

We present a novel marketing method for consumer trend detection from online user generated content, which is motivated by the gap identified in the market research literature. The existing approaches for trend analysis generally base on rating of trends by industry experts through survey questionnaires, interviews, or similar. These methods proved to be inherently costly and often suffer from bias. Our approach is based on the use of information extraction techniques for identification of trends in large aggregations of social media data. It is cost-effective method that reduces the possibility of errors associated with the design of the sample and the research instrument. The effectiveness of the approach is demonstrated in the experiment performed on restaurant review data. The accuracy of the results is at the level of current approaches for both, information extraction and market research.

LIST OF ABBREVIATIONS USED

AAFC	Agriculture and Agri-Food Canada
API	Application Programming Interface
CRFA	Canadian Restaurant and Foodservices Association
GATE	General Architecture for Text Engineering
HSX	Hollywood Stock Exchange
IE	Information Extraction
MR	Market Research
MUC	Message Understanding Conference
NLP	Natural Language Processing
OWL	Web Ontology Language
POS	Part-of-Speech
TA	TripAdvisor
UGC	User Generated Content

ACKNOWLEDGEMENTS

With this master's thesis I would like to thank the Faculties of Computer Science, Law and Management at Dalhousie University, Halifax.

I would like to express gratitude to the supervisor of the study, Dr. Vlado Keselj, because of his input, help and support through the process. I would also like to direct my appreciation to the defence committee for taking time to read my thesis and give their valuable comments. Additionally, I would like to thank all the members of Dalhousie Natural Language Processing group who shared valuable information and constructive criticism to improve this research.

CHAPTER 1 INTRODUCTION

Market research (MR) refers to any effort to gather information about markets or customers (McQuarrie, 2006). Market researchers are challenged to constantly watch for trends in the marketplace and the industry environment. Changes in the market are important because they often lead to new opportunities and threats for the company. Some examples may include changes in price sensitivity, demand for variety, level of emphasis on nutritional content, etc. Knowledge of consumer opinions and needs gained from market analysis and research is a base for business strategy (Kao, 1989). It provides objective data for decision making and validates the proposed business or product idea by confirming an existing customer demand and unfilled market niche. It may also serve as an inspiration or a starting point for product development team.

Typically companies conduct market research by collecting primary information directly from the source through focus groups, surveys, interviews, and secondary information available through credible sources such as governmental resources, industry publications, associations, etc. (Entrepreneur, 2010). However, there are several disadvantages to this standard approach. Firstly, traditional market research is based on selecting respondents and asking them sets of questions. The respondents who agree to do surveys might not be typical of the broader customer base. For example, they might be more satisfied with the product. Moreover, usually few people are willing to be a subject of the study, thus sample data may not be a representative of the target population (McGivern, 2009). Secondly, the questions asked by organizations represent their own agenda; they do not necessarily reveal the topics that consumers find important (Poynter, 2010). Thirdly, standard methods are very time consuming. Researcher needs to develop questions, find

respondents and arrange appointments with them. The whole process may be very lengthy and when completed the information may already be dated. Finally, conduction of research by using these methods may be costly. Some of these methods, such as focus groups, require a professional facilitator. Supplementary costs may include remuneration of the participants of survey, rent of a room for observations, etc.

Recently, the emergence of Internet-based social media has provided an opportunity to start a new kind of conversation among consumers and companies, creating new opportunities for organizations to understand consumers and connect with them instantly (Harvard Business Review Analytic Services, 2010). Laudon and Guercio Traver (2010) describe user content generation and social networking as a feature of e-commerce technology that unabled merchands to know much more about consumers and to be able to use this information effectively. In scholarly studies that consider how user generated content may benefit firms, the majority have focused on its use for consumer engagement and product promotion. Little is known about how the information generated by social media users can be leveraged for market research. In this study, this gap is addressed by showing that extracting and analyzing knowledge accumulated in the vast store of user generated content can yield tangible and actionable information for business and entrepreneurs.

1.1 RESEARCH QUESTION AND OBJECTIVE

The main objective of this research is to provide a methodology designed to help entrepreneurs and business professionals to extract market dynamics and identify consumer trends from large scale aggregations of online content, such as review websites.

The research question is expressed as follows:

Are results of market research based on information extraction from a large sample of user generated content comparable to traditional market research methods in identifying market trends?

The framework of this research lies in the field of electronic commerce and combines business and technology issues. It discusses avenues for business research and information extraction in e-commerce. We focus on user generated content retrieval, natural language processing, semantic analysis, and market trend analysis to develop our methodology for market research.

The significance of this study is two-fold. From the research prospective, our method proposes a systematic approach to analyze online reviews and extract knowledge. From practical prospective, it presents a cost-effective approach for companies to discover consumer knowledge, which can guide decision making, product development strategies and marketing promotion efforts. Moreover, in contrast with traditional marketing methods, our review mining method is based on retrieving what a large number of real consumers have to say about the product or service in a timely manner.

1.2 STRUCTURE OF THE DOCUMENT

The rest of this paper is organized as follows. Chapter 2 examines background and related work. Chapter 3 explains in detail the proposed method of application of online user generated content to market research. Chapter 4 presents the design of experiment that illustrates the use of presented methodology. Chapter 5 is for analysis of the results

and discussion. Finally, Chapter 6 concludes the paper and summarizes possible extensions for future work.

CHAPTER 2 BACKGROUND AND RELATED WORK

2.1 CONSUMER TRENDS

Business dictionary defines consumer trends as habits or behaviors currently prevalent among consumers of goods or services. Consumer trends track more than simply what people buy and how much they spend. Data collected on trends may also include information such as how consumers use a product and how they communicate about a brand with their social network (WebFinance Inc, 2012).

The relationship between the organizations and the people they are targeting is always dynamic (Keegan, 2009). Consumer lifestyles are a moving target. People's priorities and preferences are constantly changing. The market is evolving, competitors launch their products, brands or advertising, and consumer attitudes and behaviors change, creating new market trends. Decision makers are forced to adapt to ever changing and transforming business environment with increasing frequency. That is, whenever markets change character, or economic conditions fluctuate, or competition intensifies, or technology evolves, the need for market research becomes more critical. It is essential for companies to track consumer trends and try to anticipate them (Solomon, Zaichkowky, & Polegato, 2008). Understanding of consumer trends enables companies to keep abreast of their customers.

Stages of research process are presented in the Figure 1. Marketing research process begins with research objectives. They are the goals to be achieved by conducting research. Exploratory research helps to identify and clarify the decisions that need to be made. Exploratory research techniques include examination of previous research on the

subject, conduction of pilot study, intended to assist in design of a larger detailed study, and experience surveys. Based on the results of exploratory study, or without it, the research objectives are stated and the main research method is selected. Four basic techniques of research include surveys, experiments, secondary data and observation. These are main techniques used by researchers in the industry. A sample of respondents from target population is selected for data gathering. Data analysis includes application of statistical analysis to understand and explain the data that have been collected. Final and the most important step of analysis is the interpretation of research results and depiction of appropriate conclusions.

Standard approach to examining consumer market trends involves analysis of economic and socio-demographic data in the context of consumption to identify how changes in the economy and marketplace have affected the way consumers behave in a marketplace (Zikmund, Babin, Carr, & Griffin, 2010).

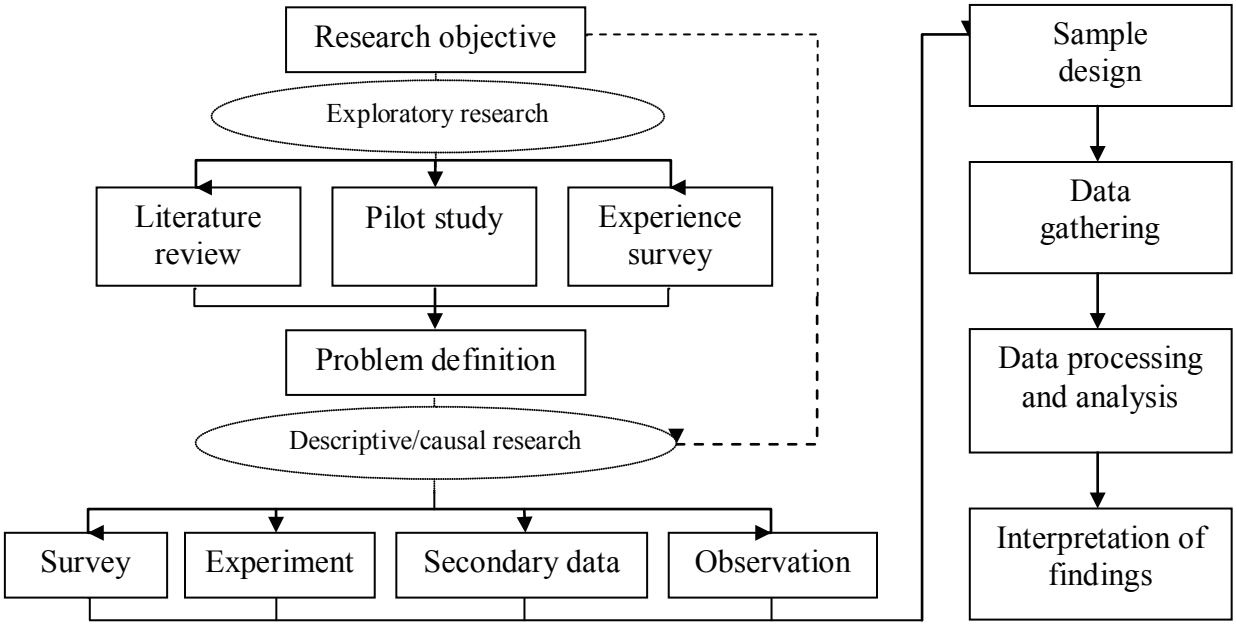

Figure 1: Flowchart of the research process (Adopted from Zikmund, et al. (2010)

In traditional approach of predicting market trends, respondents are asked to estimate how successful a concept is likely to be. The format of the interview is more like a stock market and less like a market research interview (Poynter, 2010). A sample of industry experts is selected. The industry experts are shown a list of potential trends and asked how likely they think each of these concepts is to become the next market trend. Just as stock market gambles on which companies are going to be successful, they are asked to 'gamble' on which concept or concepts they think are most likely to be successful.

An example from restaurant industry is a "Canadian chef survey" (Canadian Restaurant and Foodservice Association (CRFA), 2012). 400 restaurant chefs are asked to rank 190 menu items, ingredients, preparation methods and culinary themes into one of four categories:

- Hot trend – at a peak of popularity, customers are excited about these items and eating them more than ever;
- Up and comer – could be next hot trend, interest in these items is quickly increasing;
- Perennial favourite - popular as always, these items are long time menu favourites;
- Yesterday's news – decreased in popularity, customers are less interested in these items now.

This method assumes that professional chefs have the inside knowledge of the industry, thus they have a certain level of understanding of consumer trends. As a disadvantage of this method, it is generally known by the market researchers that expert participants may

favor their niche in the market, or have a narrow view point (McGivern, 2009). Additionally, according to neuroscience and behavioral economics, traditional market research asks questions that people may not be willing, or don't know how to answer (Poynter, 2010). Increasingly, we are beginning to realize, that market research respondents are unreliable witnesses about themselves.

A more recent example of method could be an idea proposed by James Surowiecki. He argues that the large groups of people are smarter than elite few and better at slowing problems, forecasting, innovation, coming to wise decisions, and even predicting the future (Surowiecki, 2005). An example of this idea at work could be a Hollywood Stock Exchange (HSX), which uses the technique similar to those used in forecasting for financial area. It has good reputation in predicting movie success (Levmore, 2003).

Our method represents a radical departure from the traditional paradigm of market research and offers a very different methodology for identifying new trends. Similar to Surowiecki's idea we believe that data from a large number of consumers is able to identify trends as good, or even better than few industry experts. However, the fundamental distinction of the method presented here from available marketing prediction methods is that it relies on the data generated by consumer in a free form and without an external stimulus.

It advocates that changes in consumer trends are not simply a result of demographic and economic factors. Consumers are a diverse group and it is difficult to make generalization of them (Poynter, 2010). Their needs are often dramatically different, depending not only on their age, gender, education, income, but also social circumstances, and interaction

with their environment. Limited ability of the traditional approaches for generating consumer insides and abundant information about products and services created by consumers justifies the need for a new methodology. Getting a specific understanding of consumer trends in a timely manner is the major challenge of this study.

2.2 TREND MONITORING IN THE RESTAURANT INDUSTRY DOMAIN

Trend monitoring is a generic domain into which a great deal of effort in terms of knowledge management has been placed, because almost every company, organization and business unit must encounter it. The foodservice domain is chosen for the experiment because it contains many generic kinds of concepts. It does not require a domain expert to understand the terms and concepts involved, so the system can easily be created by a person without special domain skills. This consideration is very important for fast development of a system to be used as an example application. The trend monitoring application in restaurant industry aims to alert researchers to market changes, since online reviews are a very good indicator of moving trends in the field. By monitoring restaurant reviews over a period of month and years, we can examine changes in the demand for particular menu items, interest in particular culinary themes, what kind of ethnic cuisines are being popular, etc. Application of our model for the foodservice domain is aimed at helping companies to access and monitor such information quickly and accurately, bringing new tendencies, processes and movements to their attention.

2.2.1 A SNAPSHOT OF RESTAURANT INDUSTRY DYNAMICS

Restaurant is a high-concept business. A new restaurant becomes popular when it has captured the intangible combination of food, décor, atmosphere, and unique identity that warrants the very large premium consumers pay (McCracken, 2006). It is also a high-risk

business. The difference between nailing the next trend and missing it even fractionally is the difference between a making a fortune and closing up (McCracken, 2006). The customer tolerance here is very limited.Moreover, it is a high-investment business. Restaurateurs invest significant amounts of money in creating the ambience. Finally, this business has a brief shelf life. The life-expectancy of a successful high-end restaurant can be short, especially in large cities. Taste and preference if consumers change quickly and they migrate to ever coming competitors.

This industry demands exquisite choices, big investments, and quick decision making. New projects succeed when they attract a relatively small, but precious, segment of early adopters. The people who are a little behind the curve, the middle adopters, will hear about the new place. This is a bigger following and the real profit opportunity. After that, profits start to go down. Frequently the restaurant owners have one place opened, one place about to launch and several in their minds to keep up with changing dynamics in the market (McCracken, 2006).

Dramatic changes are occurring in the way food service establishments deliver food to consumers, largely in response to dynamic and diverse trends in Canadian lifestyles. Major changes in demographics, income distribution and labor force participation continue to dictate changes in food industry. They include a slower growth in population, greater ethnic diversity, an aging population, more women in the labor force, and slower growth in income and widening disparity in its distribution (Statistics Canada, 2012). Nova Scotia has the highest median age in the country (Statistics Canada, 2012). British Columbia has seen an ever growing number of newcomers from Asian countries (Ministry of Attorney General and Minister Responsible for Multiculturalism, 2008). The

increased labor force participation of women is one of the major social and economic phenomena (OECD Economics Department, 2004). This trend is behind the much of the rising demand for services in many industries including food service. Many consumers do not have time or skills to prepare a home meal (International Markets Bureau, 2010). Low income levels and ever increasing working hours have created the demand for fast food, take out and drive through. The value of time for many consumers is higher than extra costs of purchasing food ready to eat (Kinsey & Senauer, 1996). Another group of consumers demanding fast food are price conscious consumers with low income.

The food system is consumer driven. Restaurants receive the information about the consumers' preferences from research, this information gives them power to compete with rivals effectively and respond to consumers as quickly as possible. Understanding how consumers' food demand and eating patterns are changing can help predict the future direction of the entire industry (Kinsey & Senauer, 1996). New products that most directly respond to the desires and needs of consumers are the most successful. Reformulating recipes to respond to consumers' preferences for specific food characteristics is increasingly faster and more precise. In addition, there are growing niche markets that provide unique opportunities for new businesses. Some examples are natural/organic market niche, sports food, etc.

Overall, the entire food system is very dynamic. In large part, the changes are being driven by fundamental shifts in consumer demand, by the availability of information technology, and by the quest for profits over volume (Kinsey & Senauer, 1996). Consumer trends force businesses operating in the highly competitive industry strive to

offer new ways to deliver food, new restaurant designs and layouts, new menu items and new restaurant formats.

2.3 RESEARCH GAP

Industry Canada Research Paper (Office of Consumer Affairs, 2004) identified the need for better research and better data on the consumer and the consumer's place in the marketplace today. Despite the vast quantity of economic data available in Canada, many gaps in consumer research exist. There are some significant information sources on consumers and their activity in the marketplace, but these are often limited in what they can tell us. For example, the report "Canadian consumer: behaviour, attitudes and perceptions toward food products" (2010) published by International Markets Bureau, as well as many others draws conclusions on marketing trends from demographic and sales data. It provides a useful, but very general, indicator of Canadians' interest towards authenticity and sustainability. Moreover, the report covers a very large geographic area and includes regions with dispersed consumer cultures. As a result, by its very design, the report has limited potential to provide information on the underlying factors that shape consumers' behavior in the marketplace; for example, what kind of cuisine people eat, how do they define and understand what they eat, what exactly they order etc. Analytical work is also carried out by the research communities in academia and business as well as by various for-profit market research firms. All of these sources can provide a reliable profile of consumer market. Nevertheless it is widely acknowledged that comprehensive and timely data on the market trends in local consumer groups are scarce, hard to access and expensive (International Markets Bureau, 2010). Without good data and analysis, business community is less likely to be able to determine consumer interests. Potential for

inappropriate actions by management increases. Businesses have a direct interest in understanding not only demographics and sales dynamics, but also perceptions of marketplace by consumers, in order to successfully offer the products and services consumers need, and to avoid business strategies and policies that are likely to alienate or disappoint their customers.

The international market research community has expressed interest toward market research through social media. Cooke (2009) argued in favour of taking more non-directive, collaborative, approach to market research, rather than rely solely on traditional approach based on direct questioning. He sees the content created by users on the web as a collection of searchable archives that hold unique and rich forms of information about people.

Gane and Beer (2008) suggest that user generated content offers researchers two significant opportunities: first, open and free access to data sets from which to draw inferences, second, it may be used by researchers as a set of new interactive research technologies.

Increasingly market researchers critisize the lack of insight and overrelience on an industrialized view of research, suggesting that consumer insights that are offered in creative form online should be used for market research. This type of market research is frequently called “social media research” or “new market research” (“newMR”). Ray Poynter (2010) defines newMR as a method of obtaining market research insight from social media.

The America-based Advertising Research Foundation (ARF) is leading an industry-wide 'research transformation' initiative to foster the adoption of techniques of social media research into marketing body of knowledge (Rubinson, 2009). In the article "The shape of marketing research in 2021" (Micu, et al., 2011) project that data mining and social-media research will expand rapidly, making consumer insights management a significant corporate function.

This work is inspired by above mentioned papers and the need for new research tecniques they all identify. The methodology introduced in this study tougch on several areas of data mining. The rest of this chapter will introduce the related work for knowledge discovery and information extraction.

2.4 INFORMATION EXTRACTION

Named-entity recognition was first introduced as a sub-task in the MUC (Message Understanding Conference) (Zhao & Jin, 2009). Currently, information extraction and named-entity recognition is a hot research field.

Boufaden, et al. (2005) developed a privacy complience engine that monitors outgoing emails in an organization for violation of privacy policy of this organization. As a part of their research they implemented an information extraction system that incorporates parsing and ontology based semantic tagging. Extracted information is then passed to engine for further analysis.

Boufaden, Bengio and Lapalme (2004) presented results of statistical method for the detection of generalized named entities for automatic detection of relevant words and its annotation with concepts from ontology. Previously to this work, Boufaden (2003)

presented a method for the semantic tagging of word chunks extracted from a written transcription of converstaions. Entity extraction in their approach is based on gazeteers, a domain-specific ontology and an overlapping coefficient similarity measure.

Zhao and Jin (2009) presented a system framework for the extraction of intelligence about competitor from the Web. They used two-step approach: named-entity recognition and entity relationship detection to extract company profile information(name, contact), event (time and location) and relationships (competitors and clients).

Maynard, et al. (2005) focused on enhancing gazetteer-based information extraction with ontological information. They developed an application for automatic knowledge extraction, management and monitoring in the chemical engineering domain, integrated in a dynamic knowledge management portal (h-TechSight KMP).

Baptista (2008) integrated Category and Relationship extractor (CaRE) and GATE to allow querying of medical information, such as doctor, patient, hospital, ID, location, date, and phone.

Adaptation of existing information extraction systems to new domains is the focus of much current research. One of the major bottlenecks in adapting information extraction systems to new languages is the collection and organization of new lexical resources (Maynard, Bontcheva, & Cunningham, 2004).

Maynard, Bontcheva and Cunningham in their article “Automatic language-independent induction of gazeteer lists (2004)” present a tool that collects occurences of entities directly from a small set of annotated training texts, and populates gazeteer lists with the entities.

It has been shown that a good domain specific dataset can have a vital role in adoptating existing IE systems and creating new ones. In our study the database and ontology specific to restaurant industry are presented.

Scholars have examined user-generated content as it has proliferated online.

Kivran-Swaine (2010) presented a visual analytic tool, designed to help journalists extract news value from social media data. They have shown that journalists effectively use the tool to generate insight about the social media response to the event, and about the event itself.

You, et al. (2012) focused on Chinese electronic markets and proposed a framework for extracting knowledge from online reviews through text mining and econometric analysis.

Cai, et al. (2010) presented techniques that could detect the topics associated with negative and positive opinions, which can be described as key words closely associated with each sentiment category. They identified topic words by filtering out all the stop words and opinionated words, and rated the rest by frequency and PMI value of words in each sentiment category. Frequent words with high PMI value were considered as topic words.

Despite its relevance, the issue of market research through social media has actually been scarcely researched in the literature. Thus, it seems essential to establish methodology accounting for the extraction of information from large scale user generated content on social media to analyze and validate new business ideas.

CHAPTER 3 METHODOLOGY

3.1 ARCHITECTURE OF THE SYSTEM

Market trend prediction model is composed of four components presented in Figure 2. The first component is the domain knowledge. It consists of a domain ontology describing basic types of trends, and a database containing a variety of words and word combinations that are potential consumer trends. The second component is a data collection module that collects information from the web and prepares it for the information extraction purposes. The third component is the information extraction system which extracts trend candidates from the text by using the domain ontology and a set of rules linking concepts from ontologies to classes in the database. The fourth and final component identifies market trends among extracted trend candidates. The following sections briefly describe the method design, while chapter 4 describes the use of the method for a specific dataset.

3.2 DATA COLLECTION AND CLEANING

The first step is to find information and store it in a database for analysis. Blogs, forums, Twitter, social networks, comments, anywhere where people post their opinions and advice, can serve as a source of data. An assessment determines the most suitable data source for a particular problem.

Data can be collected by use of different techniques, such as application programming interface (API), web scrapping, and automated blog search.

After the information is collected, it is examined for inconsistencies and errors using automated edits coupled with analytical review. Data cleaning process consists of

reviewing data for missing information, for clarity, error and verifying consistency between selected data fields.

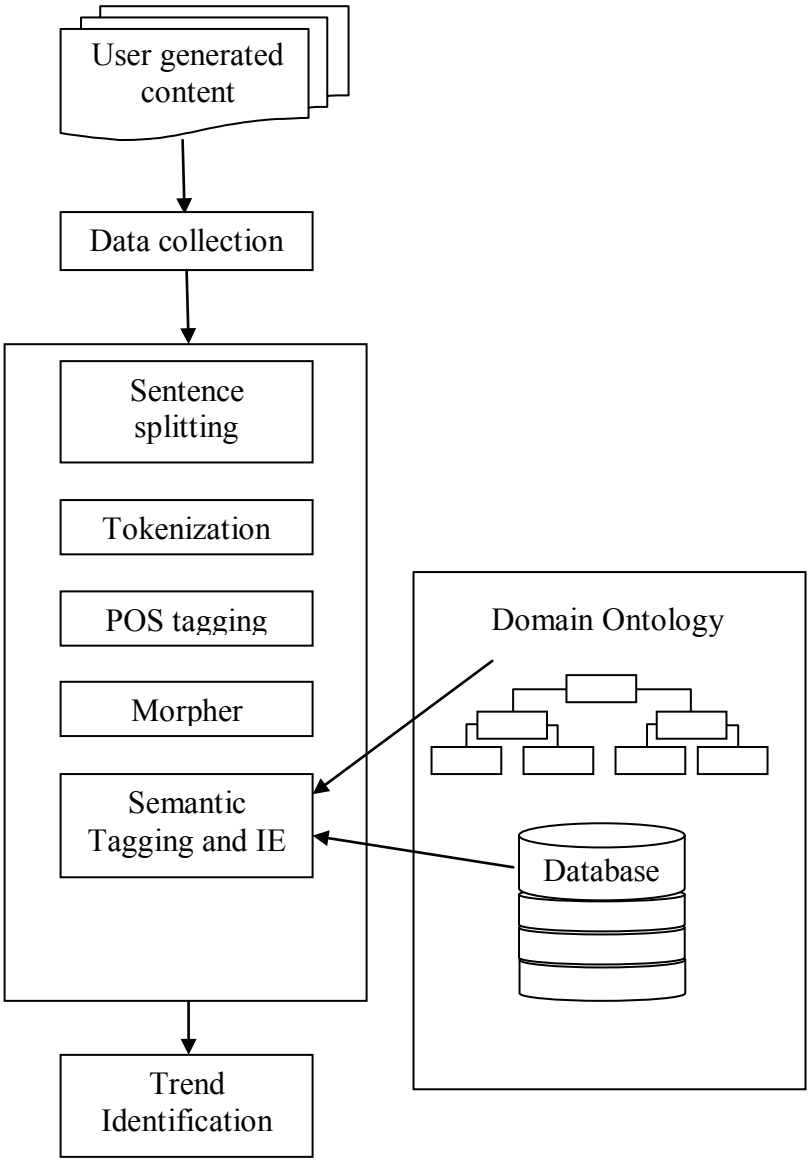

Figure 2: Architecture of the market trend prediction model

3.3 INFORMATION EXTRACTION

Information extraction is about finding and structuring relevant information in a text given a particular domain (Boufaden, et al., 2005).

We used a five-stage information extraction system that includes sentence splitting, tokenization, part-of-speech (POS) tagging, morphological analyzing, and detection of target words, its annotation and extraction developed in GATE.

3.3.1 General Architecture for Text Engineering (GATE)

GATE is an architecture for text engineering developed at the University of Sheffield (Maynard, Yankova, Kourakis, & Kokossis, 2005), containing a suite of tools for natural language processing (NLP). GATE is made up of many components that carry different functionality that can be put together to run a certain task. These components can be modified to fit needs of the particular project. Since our information extraction system is made up of several different tasks, GATE's ability to mix and match its components provides a suitable flexibility and functionality that meets our requirements. GATE is versatile. It works on different operating systems and can process other languages besides English. Finally, it is very popular open-source tool available online.

We ran GATE over corpus to produce a set of annotated texts. Initial processing stages include, among others, sentence splitting, word tokenization, part-of-speech tagging, and morphological analysis. The main processing is carried out by an ontological gazetteer and a set of grammar rules. Ontological gazetteer is a database of instances, gathered in lists. Instances are textual fragments of interest, such as words and phrases representing cuisines, food ingredients, beverages, flavours, etc. Instances are linked to the ontology by a map of rules. Ontology is a hierarchical organization that captures the subset/superset relations among the classes in the database (Jurafsky & Martin, 2009). The GATE application performs targeted information extraction relative to a domain and

ontology. Instances drawn from gathered information pave the way for monitoring trends of new and existing concepts and instances.

3.3.2 ONTOLOGY

The ontology needs to be created for a domain and populated with instances. The ontologies allow instances to generate more complex facts. Ontology consists of hierarchies of classes, instances and properties. Class hierarchy, or taxonomy, is the main component of ontology. Classes are linked relationally. Each class contains a list of instances. The task of ontology-based information extraction system is to identify instances from the ontology in the text and link them to classes.

Maynard et al (2005) advocates the use of ontology-based information extraction rather than traditional information extraction, because it implements the use of formal ontology rather than a flat lexicon or flat gazetteer lists. Ontology-based information extraction also involves reasoning, and the ability to link instances to its semantic description in the ontology.

In our approach, we encoded the domain knowledge in ontology for two reasons:

- ontologies define explicit hierarchical relations that can be used to generalize word classes and reduce their number to enhance the information extraction pattern learning process;
- they provide a grouping of word senses (meanings), so that the sense disambiguation can be reduced.

3.3.3 PROTÉGÉ OWL ONTOLOGY EDITOR

Protégé is an open source ontology development platform with functionality of editing classes, properties and instances developed at Stanford Medical Informatics with support of users. We used Protégé Web Ontology Language (OWL) plug-in. Its user interface provides various default tabs. The Classes tab displays the ontology's class hierarchy, allows developers to create and edit classes, and displays the result of the classification. The Properties tab can be used to create and edit the relations between classes in the ontology. The Individuals tab can be used to create and edit instances.

Protégé's scalability and convenience of editing provides a suitable environment for creation of ontologies for large and complex domains. Protégé supports database storage that is capable of several million instances and provides a variety of navigation aids.

3.3.4 INFORMATION EXTRACTION SYSTEM

This section presents the overview of the information extraction system. More details are provided in section 4.4. Information extraction system starts with segmenting text into sentences. Sentence splitting of the reviews is based on regular expressions. It determines text fragments that split the sentence, such as two consecutive new lines, and text fragments that seen as sentence splits, but are not, such as full stops occurring inside abbreviations. It handles a period in abbreviations, personal titles and numbers correctly.

The second stage is tokenization. At this stage sentences are further segmented into simple tokens, such as numbers, symbols, punctuation, space and words. The tokenizer is set to recognize and unite certain types of tokens. Figure 3 shows examples of defined token constructions that would be joined together.

“ ‘ “, “re” -> “ ‘re “

“90”, “s” -> “90s”

“didn”, “ ‘ “, “t” -> “did” “n’t”

Figure 3: Token constructions

The tokenizer not only splits the text into tokens, but also extracts a multitude of information on each token. Information mined during tokenization includes length of the token and orthographical information such as capitalization (Figure 4).

Pork {kind=word, length=4, orth=upperInitial}

Figure 4: Example of information extracted during tokenization

The third stage produces a part-of-speech (POS) tag as an annotation on each word or symbol. Part-of-speech tagging is the process of assigning a part of speech tag to each word in a corpus (Jurafsky & Martin, 2009). Input to POS tagger is a string of words and the tagging information created by tokenizer. The output is a single best tag for each word. GATE’s default lexicon and ruleset were used at this stage.

At the fourth stage, morphology analysis is performed. Based on part-of-speech tag and token information lemmas and affixes are identified and a root of the word assigned to each token. Morphological analysis is based on regular expression rules. These basic rules are predefined in GATE, with an option of modification by the user.

At the fifth stage, semantic tagging, the ontology and the database are used to find target words. When the root of the word found in the text matches one of the database entries, a

semantic tag is added to it. Each tag also holds a ”majorType” and “minorType” feature, which equals the instance class such as menu category or a country. This stage produces annotations to text in relation to the specific concepts in the ontology. It annotates keywords like verbs such as “grill”, “smoked”, “mashed” and nouns such as “fish and chips”, “potato”, “pork” and context of relevant information which are in this case properties of the relation such as “potato” *cooking_method* “mashed”.

The output of the system is a review text with a set of added semantic annotations. Each of the annotated words and word combinations extracted from text represent a potential market trend.

3.3.5 TREND IDENTIFICATION SYSTEM

Business research data tends to be voluminous. Statistical methods help marketers to summarize efficiently large amounts of data. In our research we use statistics techniques, frequently used in marketing (Chakrapani, 2004). Mainly, we are interested in knowing whether obvious market trends in demand for different menu segments can be detected by analyzing information extracted from customer reviews.

This stage is divided into two parts. First part analyzes extracted candidate trends identified using descriptive statistics. The second part deals with identifying the trends over time period.In the first part we use univariate descriptive analysis. It is a fairly simple but very useful and informative type of analysis, the purpose of which is to introduce the data. It is a type of an analysis that describes one variable at a time. In essence it involves summarizing and describing the data using frequency counts and frequency distributions.

A frequency count is a count of the number of times a value occurs in the dataset (McGivern, 2009). In order to detect trends in each category, we identify which of the potential trends we extracted at the previous step where mentioned by reviewers most frequently. A frequency count of the number of reviews containing each potential trend can answer this question.

CHAPTER 4 EXPERIMENT DESIGN

4.1 THE DATA

There are two steps that are essential in data collection process, which are:

1. Finding information;
2. Extracting information found and storing it in an accessible form.

The first step in the data collection process is to search for sources of user generated content (UGC) about the topic of interest on the Web.

We choose online restaurant reviews as the unit of our study. The following criteria are applied to the collected units:

1. Industry sector – food services or drinking places, including restaurants, cafes, fast food operations and other types of businesses;
2. Geographic area – the review must describe a foodservice operation in Canada.

To design a dataset that is representative of the population, we collect the reviews from different regions of the country, including both urban and rural locations. The reviews are randomly collected from a review website. The size of the sample is an important consideration. Our main purpose is to collect the largest amount of historical data possible.

Reviews were collected from TripAdvisor (TA). TA review website was chosen as a large source of reviews created by users. It is one of the world's largest travel sites featuring reviews on restaurant and other travel related services. The site has over 56

million unique monthly visitors and over 60 million reviews (TripAdvisor LLC , 2012). TA database includes 858000 restaurants around the world (TripAdvisor LLC , 2012).

Restaurant review falls into the category of unstructured text which is neither strictly formatted nor always composed of grammatical sentences. To some extent, it is similar to manually transcribed spontaneous speech. There is not always explicit punctuation, and it often contains misspellings. An example of restaurant review is given in Figure 5.

"Try the Chicken and Ribs!"

December, 2011

Zee's grill serves up fresh homemade food for a moderate price. The chicken and ribs is a fun play on a classic. I almost ordered it two days in a row. If you like real fries then this is the place. The lobster poutine is an appetizer to share. It is rich and a delicious treat. We finished up with the chocolate dessert plate. The white chocolate creme brulee is the best I have ever had. In fact, I think it should be its own dessert unencumbered by the other chocolatey offerings. We ate about three meals at Zee's during a three day stay because everything was so delicious.

Visited August 2011

Figure 5: A restaurant review

Table 1 shows some statistics on the restaurant review corpus we are working with. Our corpus is composed of 104 668 reviews of restaurants in Canada. The number of reviews is uneven in each year due to a number of reasons. First, it appears that the website might be removing old reviews. Second, the domain of online reviews is a new phenomenon,

and people are still getting used to it. However, as we can see the number of reviews is growing.

By sentiment orientation		By year of publication	
Positive reviews	66108	2006	8
Negative reviews	17111	2007	9613
Neutral reviews	21449	2008	15940
By word count		2009	7832
Total number of words	7119697	2010	11799
Average review length, words	71	2011	37360
Maximum review length, words	1524	2012	22116
Minimum review length, words	1	Total	104668

Table 1: Corpus statistics

Third, in restaurant industry businesses exit the market, move and change their names frequently. This would mean the restaurant will be deleted from the database and a new venture will appear. It therefore makes it challenging to collect all the reviews for periods distanced in time. Finally, the macro factors influence the number of reviews in each year. For instance, restaurant industry had suffered a decline in sales in 2008-2010 (Figure 6).

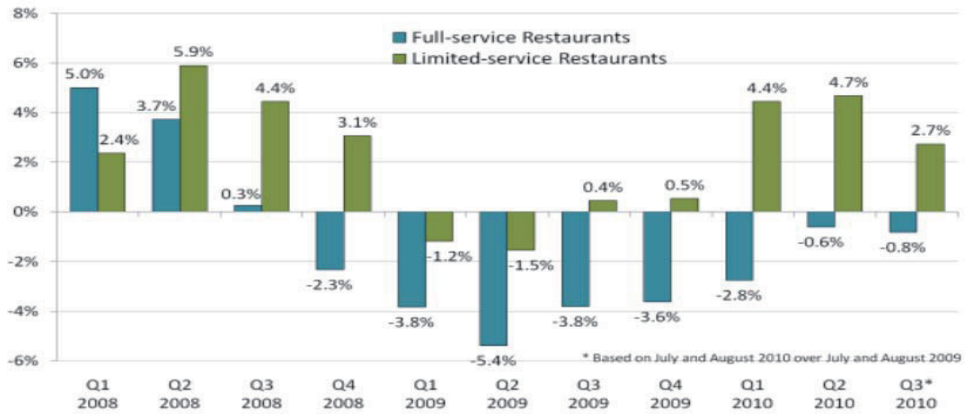

Figure 6: Restaurant Sales by Segment in Canada (Canadian Restaurant and Foodservices Association (CRFA), 2010)

Due to crises many people had limited their dining outside, resulting in negative sales growth of 5.4% for Q2 2009. It is reflected in the number of reviews (Figure 7).

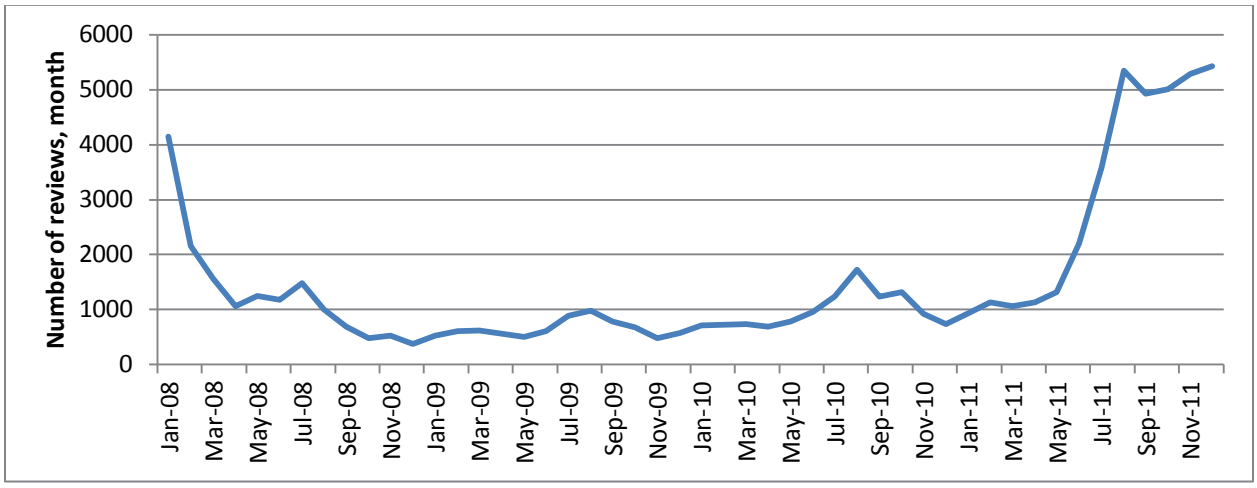

Figure 7: Number of reviews by month

Seasonality of the industry also affects the number of reviews in each month with number of reviews increasing in summer (Figure 8).

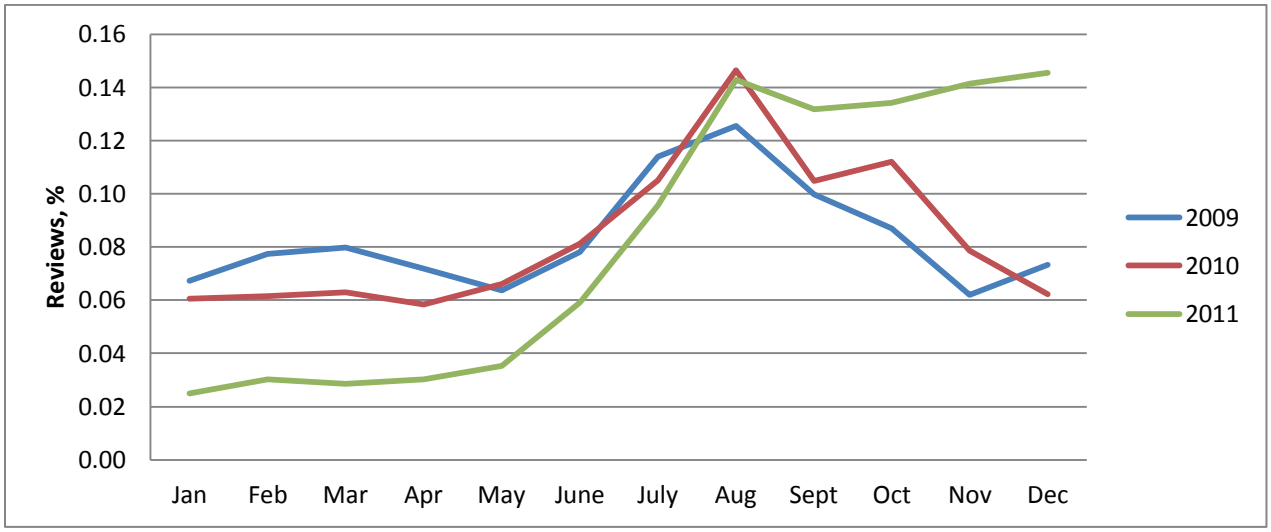

Figure 8: Seasonality of food service industry by number of reviews

4.2 DATA COLLECTION AND CLEANING

This stage is divided into three parts. The first part involves raw data collection. Reviews were crawled from the website. Since Trip Adisor limits the number of queries through API, our method incorporates automated screen scrapping program. Multiple pages of the website are viewed and information is collected directly from the source code. A page of restaurant reviews is collected and put into organized set of headings. Collected information includes review, reviewer and review target. It provides a list of entities in the following format:

```
Review(Date(Month,Year),
        Title,Text,VisitDate,
        ReviewSentOrientation,
        ReviewDetailedSent(ReviewSentFood,ReviewSentServi
        ce,ReviewSentValue,ReviewSentAtmosphere))

Reviewer(UserID,
        Location,
        TotalContributions,
        Status,
        UserVotes)

Restaurant(RestaurantName,
        Address(StreetAdd,City,Province,PostalCode,Phone)
        CuisineType,
        PriceRange,
        TotalReviews,
        RestaurantSentOrientation,
        RestaurantDetailedSent(RestaurantSentFood,Restaur
        antSentService,RestaurantSentValue,RestaurantSent
        Atmosphere)).
```

The entity `review` gives the review content as well as supplementing information whereas `reviewer` and `restaurant` gives the information about review target and review producer.

The second part deals with inconsistency and errors with the data. It is the process of checking and adjusting the data for omissions, consistency and legibility. In the experiment, the reviews in foreign languages, such as Japanese and German, and duplicate reviews were excluded from the corpus. Moreover, reviews containing only sentiment orientation information and no textual comment were omitted.

The third part addresses the segmentation of the review bodies and conversion of them into XML format. This step is an important one because information extraction component uses only content of the review, while supplementing information about the restaurant and reviewer can be used in trend analysis. The segmentation was done using XML schema to produce documents in structured form.

4.3 DOMAIN KNOWLEDGE

The ontology is built on the Canadian Restaurant and Foodservice Association's (CRFA) annual chef survey (Canadian Restaurant and Foodservice Association (CRFA), 2012). It uses knowledge about the classes defined in the survey, such as *Preparation Methods, Culinary Themes, Ethnic Cuisines and Flavours, Menu Items, Produce, Spices, Seasonings and Flavours, Alcoholic and Non-alcoholic Beverages* along with a list of instances.

We consider each instance to be a potential trend in its class. For example, an instance *organic chicken* is a potential trend in category *White Meat.* It is a part of more general trends, such as organic poultry, and organic food.

4.3.1 THE DATABASE

To represent domain knowledge about entities annotated in reviews, we build a dedicated database. We distinguish two knowledge sources for the database: International dictionary of food and cooking (Sinclair, 2005) and The new food lover's companion dictionary (Herbst & Herbst, 2007).

Foodservices in Canada provide quite extensive choice of international cuisine to its customers. There is a large number of restaurants, cafeterias and other food services that specialize in ethnic cuisine. Apart from that, the language of English cuisine is unusual in that it uses many words of foreign origin, often in their original spelling (Sinclair, 2005).Due to these two reasons an extensive amount of instances is included in the dictionary to embody international kitchen.

The dictionary includes over 35 840 terms, called instances, including ingredients, dishes, beverages, spanning over 35 languages. Terms of foreign origin, chicken tikka masala, for example, are included in their original spelling and English equivalents. The decision to include the original spelling was based on our observation of its frequent use by reviewers of restaurants serving ethnic cuisine during manual analysis of the corpus.

4.3.2 DOMAIN ONTOLOGY

The domain ontology is defined as a hierarchical organization of the main concepts described in the database (Karkaletsis, Fragkou, Petasis, & Iosif, 2011). Each concept is

represented by a class in the ontology. There is a direct mapping between classes of the domain ontology and the ones represented in the database. The ontology gives additional information about relations between database instances, e.g. “cuisine-has-menu entry”.

The ontology is used to automatically annotate parts of texts with concepts from restaurant and foodservice database. It has 14 main concepts (Figure 9): *General Menu Trends, Preparation Methods, Culinary Themes, Ethnic Cuisines and Flavours, White Meat, Red Meat/Game, Seafood/Fish, Sides, Appetizers/Starters, Desserts, Produce, Spices/Seasonings/Flavours, Alcoholic and Non-alcoholic Beverages.* Each concept in the ontology has a set of lists in the database associated with it. In total there are around 218 domain-specific lists. The lists are quite large, most of them are around 400 to 3 000 instances.

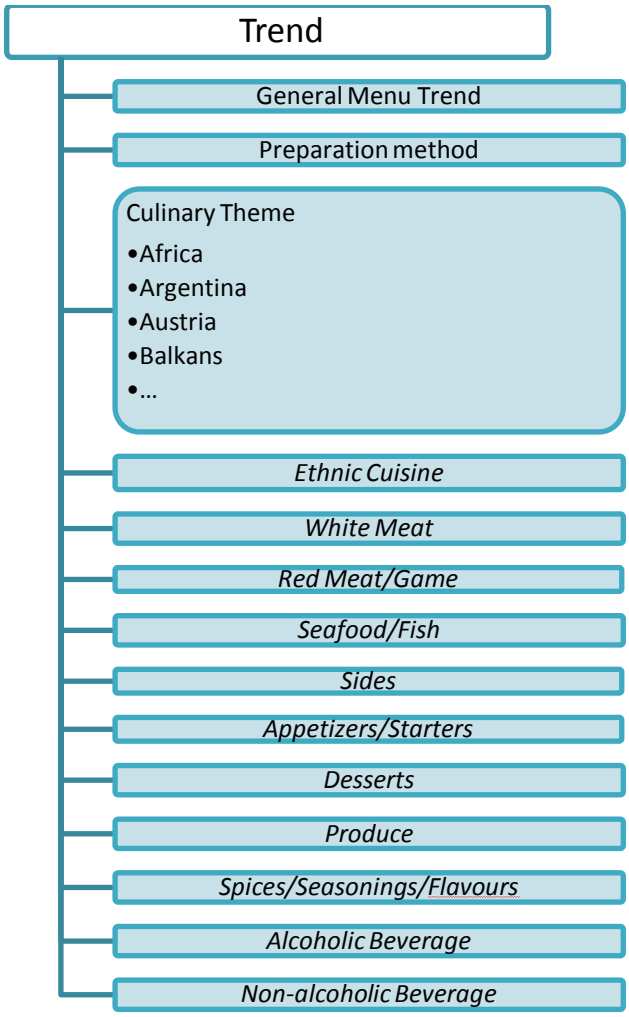

Figure 9: Domain ontology

Karkaletsis, et al. (2011) classify ontology-based information extraction in four different groups according to the level of ontological knowledge they exploit. The first level includes domain instances and their variations, as well as word classes. At a second level, domain instances or word classes are organized in conceptual hierarchy. The third level of ontological knowledge that can be exploited by IE systems concerns the word classes' properties and relations between them. The forth level of ontological knowledge is the domain model.

Our ontology is capabel of detecting named entities, according to the instance class, properties and relation types defined in the ontology. To illustrate, consider the following sentence from the experiment described in chapter 4: “I had the lunch special roast chicken penne in cream sauce with garlic toast”. First level ontology recognizes tokens “roast”, “chicken”, “penne”, “cream sauce”, “garlic”, and “toast” as named entities. Second level ontology recognizes tokens as named entities, being instances of the classes: “PreparationMethod”, “WhiteMeat”, “Cuisine” => “Italy”, “FlavourSpiceSeasoning”, “FlavourSpiceSeasoning”, and “Side”, respectively. Next, the third level ontology is capable of generating the following relations: “(roast, chicken)”, “(penne, cream sauce)” and “(toast, garlic)”. Finally, in forth level ontology, identified instances can be further exploited according to the domain model in order to build instances of higher level: “MainDish= (Dish_WhiteMeat= “Chicken”, Dish_PreparartionMethod= “roast”, Dish_Cuisine= Italy, Dish_ FlavourSpiceSeasoning= “cream sauce”, Dish_Side= “toast”, Dish_Side_ FlavourSpiceSeasoning= “garlic”)”.

Thus, with the use of the ontology based information extraction we are able to identify not only named entities mentioned in the review, but also classes they belong to, and relationships between them.

4.4 INFORMATION EXTRACTION SYSTEM

We created a domain-specific application for the foodservice field, which searches for relevant information in the text. In foodservice context, relevant information is, for example, menu entry names, ingredients, ethnic dishes, and preparation methods. These are considered to be potential trends in restaurant industry (Canadian Restaurant and Foodservice Association (CRFA), 2012).

The application is a multi-stage process of text analysis. Fist, the corpus of restaurant reviews is loaded into application. The reviews are broken into separate sentences. After that, the application takes one sentence at a time and extracts simple tokens and additional information about them, assigning POS tag to word tokens. Based on the information obtained, the basic form of the word, the root, is identified, followed by a database look up. If the word or word combination is listed in the database, the application assigns a tag to it. The tag contains information about the class and sub-class of the word in the database. The output of the last stage is presented in Figure 10.

The application was able to identify 5 potential trends in the review text: “Thai”, “black pepper”, “pork”, “seafood”, and “pad”. For example, when the application picked up the fragment “Thai”, it recognized it as a word, produced its length and recorded that the first letter in this fragment is spelled in upper case. Further, it was identified as noun phrase and the initial form of the word “thai” was assigned. Due to the fact that Thai cuisine is considered to be a potential trend in category “ethnic cuines” and the word “thai” is listed in the sub-category “country”, the word “thai” is assigned two tags (“ethnic cuisines” and “country”) and extracted.

Review:

2007

Best Viet-Thai restaurant in the area. Extensive menu catering to both individual eating or sharing. Best visited with 7 people sharing 8 dishes. Try the (421) Black Pepper Pork and (420) Seafood Pad Thai. Delicious. Oh yeah – service is not great, but who cares when the food is this tasty.

Thai	{maorType=ethnic_cuisines, minorType=country}	{category=NNP, kind=word, length=4, orth=upperInitial, root=thai, string=Thai}
Black Pepper	{maorType=spices_seasonings_flavours}	{category=NNP, kind=word, length=12, orth=upperInitial, root=black pepper, string=Black Pepper}
Pork	{maorType=red_meat}	{category=NNP, kind=word, length=4, orth=upperInitial, root=pork, string=Pork}
Seafood	{maorType=fish_and_seafood}	{category=NNP, kind=word, length=7, orth=upperInitial, root=seafood, string=Seafood}
Pad	{maorType=ethnic_cuisines, minorType=Thailand}	{category=NNP, kind=word, length=3, orth=upperInitial, root=pad, string=Pad}

Figure 10: Information Extraction system's output

4.5 TREND IDENTIFICATION SYSTEM

We computed a frequency count for each feature extracted by our application in all years to get a glance of the trends before preparing a detailed analysis. It allowed us to see the size of each sub-group of trends, what categories might be grouped together, and what weighting might be required. The information extraction system found over 31181 reviews containing potential trends. Table 2 shows the number of reviews in each category, and the number of potential trends.

Category	Number of reviews	Number of potential trends
Menu Trends	3207	23
Preparation methods	16391	172
Culinary Themes	2492	33
Ethnic Cuisines and Flavours	37369	1488
White Meat	3508	142
Red Meat/Game	10275	228
Seafood/Fish	8458	217
Sides	12803	52
Appetizers	2108	90
Desserts	5441	185
Produce	8586	266
Spices, Seasonings, and Flavours	7590	130
Non-alcoholic Beverages	4468	110
Alcoholic Beverages/ Cocktails	6080	355

Table 2: Number of reviews and features in each category

As we can see in the table above, the number of potential trends varies greatly in each category. We define the trend to be one of the top 20 most frequent features in each

category. The frequency count expressed in raw numbers was reduced to percentages to provide a comparison of the data between trends, to see relative size of each trend and a group of trends.

As an example, the following trends were identified in the category “Spices, seasonings and flavours” (Figure 11). It can be seen from figure that the popularity of spicy, salty and BBQ flavours in food changes over time, with spicy food being the most popular. Comparable tendencies can be observed from CRFA’s reports (Figure 12).

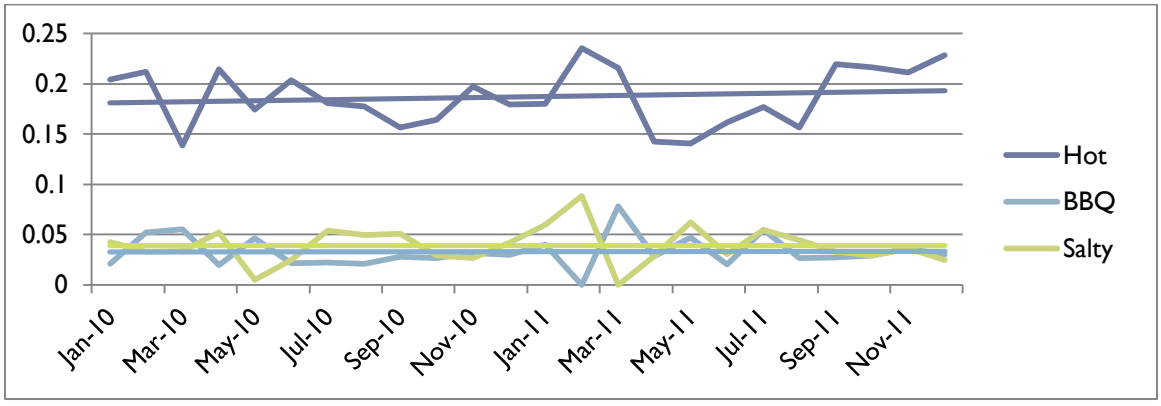

Figure 11: Top trends in category "Spices, seasonings, and flavours"

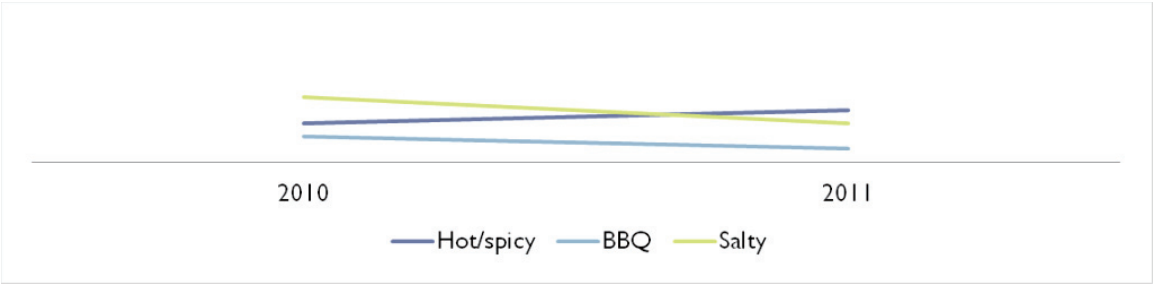

Figure 12: Top trends in category "Spices, seasoning, and flavours by CRFA

CHAPTER 5 RESULTS AND DISCUSSION

5.1 EVALUATION OF IE SYSTEM

Evaluation is carried out by comparing output of the system with manually human annotated reviews. 1200 reviews were used for manual annotation. Out of these 559 reviews contained potential trends. The total 977 tags were assigned. Information extraction system picked up 496 reviews and assigned 793 semantic tags correctly.

Table 3 shows the recall and precision scores for the information extraction system. Precision is the number of relevant documents a search retrieves divided by the total number of documents retrieved, while recall is the number of relevant documents retrieved divided by the total number of existing relevant documents that should have been retrieved. The F measure provides a way to combine these two measures into a single metric. The F measure is defined as $F_1 = 2PR/P + R$.

It is clear that semantic annotator misses out 17% of the potential trends. A significant amount of errors is related to typos, misspellings and grammatical errors. Moreover, unstructured nature of online reviews decreases the part-of-speech tagging performance which results in increasing the semantic labelling error.

	Recall	Precision	F measure
IE system	83%	96%	89%

Table 3: Recall, precision and F score of IE system on 1200 reviews

All the errors in precision are attributed to homonyms, i.e. words and word phrases with identical spelling, but different meaning, and incorrectly attributed adjectives. For

example, in the sentence "I had a pair of Chinese friends to join me for dinner", the reviewer is talking about people, but the system assumes that he is describing Chinese cuisine.

According to Jurafsky and Martin (2009), high-performing systems achive entity level F measures around 92% for PERSONS and LOCATIONS, and around 84% for ORGANIZATIONS. Considering the fact we conducted the experiments on flexible free form text, rather than well written English texts, we consider our results comparable to those of other similar research.

5.2 Evaluation of trend identification system

The major result is assessment of the feasibility of trend identification by the system. It was carried out by comparing the output of the system with official industry reports, such as annual report of Canadian Restaurant and Foodservice Association (CRFA) and marketing analysis report by Agriculture and Agri-Food Canada (AAFC).

5.2.1 Evaluation of results in comparison with CRFA's 2010 survey

It is conducted in cooperation with Canadian Culinary Federation and considered to be one of the largest research in predicting trends in restaurant industry (Eliott, 2011).

The survey provides a list of top 5 trends in each category of menu items and cooking methods. We computed frequency measures for trends identified by IE system and used it for comparison with hot trends, described in the report as menu items and cooking methods that are at the pick of popularity; customers are eating them more than ever. We allowed a degree of flexibility by compearing top 20 of our findings versus top 5 of CRFA's, in view of the fact that the results of the studies are based on different kinds of

data sets. As it can be seen from an outcome presented in Table 4 results of our analysis correlate with the findings of CRFA's chef survey.

The method allowed us to directly identify 53% of trends recognised by CRFA's report in 2010. This result is considered to be significant, based on the significant difference in the data sources and research methodology.

However, we believe that review writers may not be aware of the origins of the food ingredients. This includes the specifics of preparation methods, such as cooking with tea, or liquid nitrogen chilling. People may not be able to recall, spell properly the regional ethnic cuisine, or opt out for a simpler way by mentioning the country of origin of the food they have ordered. They may not be aware if they are eating organic, grass-fed or free-range poultry, unless the food service provider has informed them about it. We assume that these trends are a part of more general trend identified by the system. For example, the trend "free-range poultry" is a hidden trend of the category "poultry". After adjusting the results with the assumption that a group of trends is equally important as a trend we obtained 75% of trends correlated with the trends identified by CRFA's survey.

5.2.2 Evaluation of results in comparison with AAFC's 2010 report

Additionally, we compared the trends identified by our method with the findings of market analysis report, conducted by Agriculture and Agri-Food Canada in May of 2010. The report provides information on Canadian consumers, highlighting the demographics, behaviours and attitudes that influence their demand for food and beverage products (International Markets Bureau, 2010). They identify the two categories of trends related to research: long-established consumer trends and emerging trends.

Trend	Our approach	CRFA's survey	Trend	Our approach	CRFA's survey
General Menu Trends			**Sides**		
Small plates	●	●	Artisanal cheeses		●
Inexpensive	●	●	Ethnic condiments		●
Half portions	●	●	Flatbreads		●
Gourmet sandwiches	●	●	Ancient grains		●
Gourmet burgers	●	●	Charcuterie		●
Preparation Methods			**Appetizers/Starters**		
Sous vide		●	Mini-burgers		●
Liquid nitrogen chilling		●	Amuse bouche	●	●
Smoking	●	●	Appetizer combos	●	●
Braising		●	Edamame	●	●
Cooking with tea		●	Asian appetizers	●	●
Culinary Themes			**Desserts**		
Locavore	●	●	Locally grown fruits	●	●
Sustainability	●	●	Bite-size desserts		●
Farm-to-fork	●	●	Dessert platters		●
Simplicity	●	●	Sweet and salty desserts	●	●
Nutrition / health	●	●	Drinkable desserts		●
Ethnic Cuisines			**Produce**		
Ethnic fusion		●	Locally grown produce	●	●
Regional ethnic cuisine		●	Organic produce		●
Southeast Asian	●	●	Heirloom tomatoes		●
Indian	●	●	Superfruits		●
Middle Eastern	●	●	Specialty potatoes	●	●
White Meat			**Spices, Seasonings and Flavours**		
Organic poultry		●	Salt (flavored, smoked)	●	●
Free-range poultry/ pork		●	Marinades / rubs		●
Pork belly	●	●	Hot / Spicy flavours	●	●
Pork	●	●	Bold BBQ flavours	●	●
Guinea fowl		●	Foams	●	●
Red Meat/Game			**Non-alcoholic Beverages**		
Locally produced red meat		●	Organic/ Fair-trade coffee	●	●
Grass-fed beef		●	Green tea	●	●
New cuts of meat	●	●	Energy drinks		●
Aged meats		●	Specialty iced tea	●	●
Game meats	●	●	Flavoured water		●
Seafood/Fish			**Alcoholic Beverages/Cocktails**		
Local seafood/fish	●	●	Local wine and beer	●	●
Sustainable seafood		●	Culinary cocktails	●	●
Non-traditional fish		●	Craft beer		●
Fresh local oysters		●	Bar chefs/ mixologists		●
Spot prawns	●	●	Artisan liquor		●

Table 4: Comparison of findings (2010)

The long-established trends identified by AAFC include:

- Value – products meeting the needs and/or wants of the individual, at an acceptable price (e.g. inexpensive, generous portion size);
- Health – functional foods, foods and beverages that contain specific components that provide health benefits (e.g. low trans fat/sugar/sodium/salt content, whole grains, organic);

The emerging trends identified by AAFC include:

- Authenticity – products that incorporate a mixture of attributes such as geographical provenance, ethnicity, nostalgia, or a historical or expert preparation technique (e.g. natural, organic, local, fair-trade, free-range, farm-to-fork);
- Sustainability – foods that are produced in sustainable manner (e.g. food miles, carbon footprint, locavore).

The outcomes of our research confirm the results of AFFC's report. According to our findings, health, value, authenticity and sustainability are general trends regularly talked about in reviews. 59% of reviews mentioned healthy food and ingredients, including salads, vegetables and fruits on the menu. Including 10% of people emphasized their deliberate choice of organic food and beverages, vegan, vegetarian, gluten-free entries, etc. Over 45% of reviews discussed authentic kitchen, seeking to enjoy gourmet food, food, prepared using traditional methods and flavours of international cuisines, simple and 'back to basics' cooking. About 6% of reviews discussed the combination of price versus value, such as half portion options, and large size servings. About 2% of reviews expressed their concerns about sustainability, farm-to-fork and locavore menu entrees.

5.3 EVALUATION SUMMARY

The results from this chapter provide evidence that similar results are obtained from market research method presented here and official industry studies. From these results we can agree that the results of information extraction from a large sample of user generated content are comparable to those of traditional market research methods in identifying market trends.

5.4 LIMITATIONS OF PROPOSED METHOD

We identify the following limitations of the proposed methodology:

- Precise numerical estimates are usually expected as the market research outcome. This methodology is not designed to provide numerical estimates of market size, growth rate, or product price. Market research technique presented here is not quantitative in nature. It is rather intended for qualitative market analysis.
- It can be argued that sheer complexity of technological products, and the inability of customers to articulate or even envision how they would use something that doesn't yet exist, makes market research techniques ineffectual when technological innovation is the goal. The methodology presented here is considered by us as relevant to market research aiming less radical innovation. For example, when a new, slightly modified version of already existing product is planned, the techniques presented in this paper may be applicable in some point in the process. However, market research of any kind may be premature, not cost-justified and of limited value for radical innovations, when general public is unfamiliar with the product. In this case, the method

presented in this paper will make its contribution after the innovation is introduced with the aim of maximizing the odds of product's success in the market.

- It is rarely the case that a research problem can be addressed by the application of a single technique in isolation (McQuarrie, 2006). More commonly a set of research techniques is applied in sequence. The method presented in this study is not an exception. We suggest it can be used independently or in conjunction with other research methods. For example, it can serve as the basis of creating the survey questionnaire.
- According to Statistics Canada, the age is significantly correlated with Internet use. The differences in Internet usage are reflected in statistics on social media, which show that younger groups use these modes of communication to a far greater extent than older groups (Dewing, 2010). In 2007, 34% of Canadians aged 16 to 34 contributed content on the Internet, while only 13% of users aged 35 to 54, and 13% of those aged 54 and older (Dewing, 2010). These statistics should be taken into account, when projecting the findings onto the general population.

CHAPTER 6 CONCLUSION AND FUTURE WORK

Understanding of the industry trends is critical for companies in improving their performance and keeping up with competitors. Thanks to the large amount of reviews, feedback and comments from online consumers on the products this information can be easily collected and analyzed. This study utilized the available technologies in text mining and natural language processing to propose a method for businesses to analyze market trends and create business intelligence with low cost.

The main objective of this research was to develop a methodology intended to help entrepreneurs and business professionals to extract consumer dynamics and identify market trends from large scale aggregations of online content, such as review websites. The results indicated that the method developed here was able to generate results closely correlated with those of traditional methods but this comparison is not easily tested. The F measure reported in the previous section provides support that the method developed is among industry's high standard.

Our experiment on analyzing reviews shows that the methodology we proposed is feasible. Compared to questionnaires, the traditional method to study industry trends, the approach we proposed is more cost effective and superior in timely identifying novel tendencies that might not be noticed by the researchers otherwise. The result could serve as a reference for questionnaire design. With text mining, the framework has great potential for applications in different industries.

Although reviews in the experiment are all from one website, one can apply the same approach using reviews from different websites or product review oriented blogs. The

methodology we proposed can be extended to reviews of different languages by adapting the feature extraction step.

This study contributes to social media mining literature in several ways. First, it improves the understanding on how social media can be used to extract knowledge about consumer market. This leads to a more differentiated view on how analysis of user generated content contributes to the market research and business decision making. It also gives more precise guidance to managers who want to make the best possible use of convincing power of social media. In addition, the findings add to the literature on market research in showing that the social media can be leveraged for business use to support industry trend identification and product development.

The rest of this chapter discusses the future work.

6.1 LACK OF METHODOLOGY FOR MARKET RESEARCH THROUGH SOCIAL MEDIA

Looking ahead it would be advisable for researchers from government, business, academia and the non-governmental organization community to create new analytical frameworks that utilize online content created by users in their future work in the area of market research.

Every organization faces difficult choices about where to devote limited resources. All actors have to make these choices based on their own capabilities, resources, and responsibilities or interests. Clearly, having a broad choice of analytical methods will help to ensure that the managerial decisions are well founded. In this study we developed a methodology for consumer trend detection through social media. However, developing

effective frameworks for conducting various other market research tasks is a major issue that needs to be addressed in the future.

6.2 Consumer trends forecasting

Using the methodology presented above one could identify and monitor consumer trends in the industry of interest. The velocity of social media allows doing it in a dynamic fashion. Providing a decision maker with business intelligence in real time instead of data collected 3 month ago allows quick course corrections and better decisions. The fresher the data, the more valuable it is to an organization.

The next logical step is to use it to build the model for trend forecasting. The outcome of our experiment is a dataset of yearly trends. It can be used for building, training and evaluation of predictive model. In regard to our experiment, we can use the monthly data for 2007, 2008 and 2009 to build the model and apply it to predict the trends in 2011. To evaluate how good the model works, we use the data for 2010. We expect the overall model outcomes to be better than random guess.

6.3 Sentiment analysis for prediction of positive and negative consumer trends

The scope of the problem (trend identification) made sentiment analysis unnecessary. The application of the proposed methodology to other problems may require a reliable algorithm for sentiment detection. The advantageous aspect of incorporating the sentiment analysis in trend prediction practice is that the information mined can be more specific.

BIBLIOGRAPHY

1. Agresta, S., & Bough, B. B. (2011). *Perspectives on Social Marketing.* Boston: Cengage Learning.

2. Baptista, S. (2008). *Integrating medical text extraction tools.* Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology.

3. Bijmolt, T. H., S.H., L. P., Block, F., Eisenbeiss, M., Hardie, B. G., Lemmens, A., & Saffert, P. (2010, August 11). Analytics for Customer Engagement. *Journal of Service Research*.

4. Blanchard, O. (2011). *Social media ROI: Managing and Measuring Social Media Efforts in Your Organization.* Indeanapolis: Pearson Education.

5. Boufaden, N. (2003). *An ontology-based semantic tagger for IE system.* Proceedings of the ACL-2003 Student Research Workshop.

6. Boufaden, N., Bengio, Y., & Lapalme, G. (2004). *Extended semantic tagging for entity extraction.* Lisbon, Portugal: Beyond Named Entity Recognition Semantic labeling for NLP tasks Workshop held Jointly with LREC 2004.

7. Boufaden, N., Elazmeh, W., Matwin, S., Ma, Y., El-Kadri, N., & Japkowicz, N. (2005). *PEEP - An Information Extraction based approach for Privacy Protection in Email.*

8. Bygrave, W., & Zacharakis, A. (2008). *Entrepreneurship.* John Wiley & Sons, Inc.

9. Cai, K., Sprangler, S., CHen, Y., & Zhang, L. (2010). Leveraging sentiment analysis for topic detection. *Web Intelligence and Agent Systems: An International Journal*, 291-302.

10. Canadian Restaurant and Foodservice Association (CRFA). (2012). *CRFA's 2012 Canadian Chef Survey.* Canadian Restaurant and Foodservice Association (CRFA).

11. Canadian Restaurant and Foodservices Association (CRFA). (2010, Novmber). *Submission for the Pricing Consultations.* Retrieved July 5, 2012, from Canadian Restaurant and Foodservices Association (CRFA): http://www.crfa.ca/news/2010/crfa_calls_for_fairness_in_dairy_pricing_submission.pdf

12. Chakrapani, C. (2004). *Statistics in market research.* Hodder Education.

13. Claster, W., Caughron, M., & Sallis, P. (2010). Harvesting Consumer Opinion and Wine Knowledge Off the Social Media: Grape Vine Utilizing Artificial Neural Networks. *UK Sim Fourth European Medelling Symposium on Computer Modelling Simulation* (pp. 206-211). IEEE Computer Society.

14. Cooke, M. (2009). WARK: Online research. *Social media and market research: we are becoming a listening economy and, while the future of market research is bright, it will be different* (pp. 550-553). London: International journal of market research.

15. Cunningham, H., Maynard, D., & Bontcheva, K. (2011). *Text Processing with GATE (Version 6).* Gate.

16. Dewing, M. (2010). *Social Media: who uses them.* Ottawa: Library of Parliament.

17. Diakopoulos, N., Naaman, M., & Kivran-Swaine, F. (2010). Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry. *IEEE Symposium on Visual Analytics Science and Technology* (pp. 115-122). Salt Lake City: IEEE.

18. Domingos, P., Lowd, D., Kok, S., Poon, H., Richardson, M., & Singla, P. (2008). *Just Add Weights: Markov Logic for the Semantic Web.* Berlin: Springer-Verlag Berlin Heidelberg.

19. Eliott, C. (2011). 2011 Canadian Chef Survey. *Candian Restaurant and Foodservice News*.

20. Entrepreneur. (2010, September 30). *Conducting Market Research*. Retrieved November 16, 2011, from Entrepreneur: http://www.entrepreneur.com/article/printthis/217388.html

21. Fischer, E., & Reuber, R. (2010). Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behaviour? *Journal of Business Venturing*, 1-18.

22. Gane, N., & Beer, D. (2008). *New media.* Berg: Oxford.

23. Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2010). Information Diffusion Through Blogspace. *SIGKDD Explorations*, 43-52.

24. Harvard Business Review Analytic Services. (2010). *The New Conversation: Taking Social Media from Talk to Action.* Harvard Business School Publishing.

25. Heinemann, G. (2007). *Successful Entrepreneurial Market Research Techniques*. Retrieved November 20, 2011, from Entrepreneurship: http://www.entrepreneurship.org/en/resource-center/successful-entrepreneurial-market-research-techniques.aspx

26. Herbst, S. T., & Herbst, R. (2007). *The new food lover's companion more than 6,700 A-to-Z entries describe foods, cooking techniques, herbs, spices, desserts, wines, and the ingredients for pleasurable dining.* Hauppage, N.Y.: Barron's Educational Series, Inc.

27. International Markets Bureau. (2010). *The Canadian consumer: behaviour, attitudes and perceptions toward food products.* Ottawa: Agriculture and Agri-food Canada.

28. Jackson, B. (2010, August 18). *Mid-sized firms more likely to consider social media 'waste of time'.* Retrieved November 6, 2011, from itbusiness.ca: http://www.itbusiness.ca/it/client/en/home/detailnewsprint.asp?id=58817

29. Jurafsky, D., & Martin, J. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River: Pearson Education.

30. Kao, R. (1989). *Entrepreneurship and enterprise development.* Toronto: David Collinge.

31. Karkaletsis, V., Fragkou, P., Petasis, G., & Iosif, E. (2011). Ontology based information extraction from text. *Multimedia Information Extraction* (pp. 89-109). Berlin Heidelberg: Springer-Verlag.

32. Keegan, S. (2009). *Quantitative research: good decision making through understanding people, cultures and markets.* Philadelphia, USA: Kogan Page.

33. Kinsey, J., & Senauer, B. (1996). Consumer trends and changing food retailing formats. *American Agricultural Economics Association*, 1187-1191.

34. Laudon, K. C., & Guercio Traver, C. (2010). *E-Commerce: Business. Technlogy. Society.* Prentice Hall.

35. Levmore, S. (2003). Simply Efficient Markets and the Role of Regulation: Lessons from the Iowa Electronic Markets and the Hollywood Stock Exchange. *JOURNAL OF CORPORATION LAW*, 589-606.

36. Lim, S.-H., Kim, S.-W., Park, S., & Lee, J. H. (2011). Determining Content Power Users in a Blog Network: An Approach and Its Applications. *IEEE Transactions on systems, man, and cybernetics*, 853-862.

37. Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Berlin: Springer.

38. Maynard, D., Bontcheva, K., & Cunningham, H. (2004). *Automatic language-independent induction of gazeteer lists.* In Proceedings of 4th Language Resources and Evaluation Conference .

39. Maynard, D., Yankova, M., Kourakis, A., & Kokossis, A. (2005). Ontology-based information extraction for market monitoring and technology watch.

40. McCracken, G. (2006). *Flock and flow: predicting and managing change in a dynamic marketplace.* Bloomington: Indiana University Press.

41. McGivern, Y. (2009). *The practice of market research.* Harlow, England: Pearson Education Limited.

42. McQuarrie, E. F. (2006). *The market research toolbox: A concise guide for beginners.* Thousand Oaks: Sage Publications, Inc.

43. Micu, A., Dedeker, K., Lewis. (2011). The shape of marketing research in 2021. *Journal of Advertizing Research*, 213-221.

44. Ministry of Attorney General and Minister Responsible for Multiculturalism. (2008, June). *Census Fact Sheet: The Diversity of Visible Minorities and Ethnic Origins in BC.* Retrieved June 15, 2012, from www.welcomebc.ca: http://www.welcomebc.ca/local/wbc/docs/communities/visible_minorities_ethnic_origins.pdf

45. OECD Economics Department. (2004). *Female labour force participation: past trends and main determinants in OECD countries.* OECD Economics Department.

46. Office of Consumer Affairs. (2004). *The consumer trends report.* Industry Canada: Industry Canada.

47. Pang, B., & Lee, l. (2008). Opinion Mining and Sentiment Analysis. *Computational Linguistics*, 311-312.

48. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Emperical Methods in Natural Language Processing (EMNLP)* .

49. Pasupathy, K. (2010). Forecasting model for strategic and operations planning of a nonprofit health care organization. In K. Lawrence, & R. Klimberg, *Advances in Business and Management Forecasting* (pp. 59-69). Bingley: Emerald Group Publishing Limited.

50. Poynter, R. (2010). *The Handbook of Online and Social Media Research: Tools and Techniques for Market Researchers.* West Sussex: Wiley.

51. Rubinson, J. (2009). The New Marketing Research Imperative: It's about Learning. *Journal of Advertising Research*.

52. SAS Canada. (2011, August 30). *Less than a fifth of Canadian companies use social media effectively*. Retrieved November 6, 2011, from www.sas.com: http://www.sas.com

53. Sinclair, C. G. (2005). *Dictionary of Food: International Food and Cooking Terms from A to Z.* A&C Black Publishers Ltd.

54. Solomon, M. R., Zaichkowky, J. L., & Polegato, R. (2008). *Consumer behaviour: buying, having, and being.* Toronto, Canada: Pearson Education.

55. Statistics Canada. (2012, May 29). *Visual Census. 2011 Census.* Retrieved June 15, 2012, from Statistics Canada: http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/vc-rv/index.cfm?Lang=ENG&TOPIC_ID=1&GEOCODE=01

56. Surowiecki, J. (2005). *The wisdom of crowds.* New York: Anchor Books.

57. Thet, T. T., Na, J.-C., & Khoo, C. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 823-848.

58. TripAdvisor LLC . (2012). *Fact Sheet: TripAdvisor*. Retrieved July 1, 2012, from http://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html

59. WebFinance Inc. (2012). *BusinessDictionary.com*. Retrieved 07 01, 2012, from http://www.businessdictionary.com/definition/consumer-trends.html

60. Yang, C., & Ng, T. (2009). Web Opinions Analysis with Scalable Distance-Based Clustering. *IEEE Computer Society*, 65-70.

61. Yang, C., & Ng, T. (2011). Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-Based Clustering. *IEEE Transactions on Systems, man, and Cybernetics*, 1144-1155.

62. You, W., Xia, M., Liu, L., & Liu, D. (2012). Customer knowledge discovery from online reviews. *Electronic Markets*.

63. Zeng, D., Chen, H., Lush, R., & Li, S.-H. (2010). Guest Editor's Introduction. In *Social Media Analytics and Intelligence* (pp. 13-16). IEEE Intelligent Systems.

64. Zhao, J., & Jin, P. (2009). *Towards the Extraction of Intelligence about Competitor from the Web*. Springer-Verlag Berlin Heidelberg.

65. Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2010). *Business Research Methods*. Mason, OH: South-Western Cengage Learning.