Handbook of Research on Web 2.0, 3.0, and X.0:
Technologies, Business, and Social Applications

San Murugesan
Multimedia University, Malaysia & University of Western Sydney, Australia

Volume I

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey · New York

Director of Editorial Content: Kristin Klinger
Senior Managing Editor: Jamie Snavely
Assistant Managing Editor: Michael Brehm
Publishing Assistant: Sean Woznicki
Typesetter: Carole Coulson, Ricardo Mendoza, Kurt Smith
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

Library of Congress Cataloging-in-Publication Data

Handbook of research on Web 2.0, 3.0, and X.0 : technologies, business, and social applications / San Murugesan, editor.
p. cm.
Includes bibliographical references and index.
Summary: "This book provides a comprehensive reference source on next generation Web technologies and their applications"--Provided by publisher.
ISBN 978-1-60566-384-5 (hardcover) -- ISBN 978-1-60566-385-2 (ebook) 1. Web 2.0. 2. Social media. I. Murugesan, San.
TK5105.88817.H363 2010
025.042'7--dc22
2009020544

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 18
Accessing, Analyzing, and Extracting Information from User Generated Contents

Paolo Casoto
University of Udine, Italy

Antonina Dattolo
University of Udine, Italy

Paolo Omero
University of Udine, Italy

Nirmala Pudota
University of Udine, Italy

Carlo Tasso
University of Udine, Italy

ABSTRACT

The concepts of the participative Web, mass collaboration, and collective intelligence grow out of a set of Web methodologies and technologies which improve interaction with users in the development, rating, and distribution of user-generated content. UGC is one of the cornerstones of Web 2.0 and is the core concept of several different kinds of applications. UGC suggests new value chains and business models; it proposes innovative social, cultural, and economic opportunities and impacts. However, several open issues concerning semantic understanding and managing of digital information available on the Web, like information overload, heterogeneity of the available content, and effectiveness of retrieval are still unsolved. The research experiences we present in this chapter, described in literature or achieved in our research laboratory, are aimed at reducing the gap between users and information understanding, by means of collaborative and cognitive filtering, sentiment analysis, information extraction, and knowledge conceptual modeling.

DOI: 10.4018/978-1-60566-384-5.ch018

INTRODUCTION

The Web of the 1990s, identified after as *Web 1.0*, has been a read-only medium for the majority of users, even if the original idea of Tim Berners-Lee was related to a read-write Web (the first browser, named WorldWideWeb, was also a HTML editor). In 2004, the term *Web 2.0* firstly used by Dale Dougherty during a O`Reilly Media brainstorming session has been defined by Tim O`Reilly (2007) as "the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform". Web 2.0 is characterized by active participation and interaction of users that become Web's authors and can directly create, express themselves and communicate.

The innovative approach represented by Web 2.0 is only marginally related with the availability of a real technological advance in intercommunication technologies, it represents rather a new way of thinking, a new business opportunity that makes it very simple to create and share contents online and transforms every individual user of the Web into a potential producer; in this way, users may express themselves through User-Generated Content (UGC). Examples of UGC range from social bookmarking (e.g., del.icio.us) to photo and video sharing (e. g., Flickr and YouTube), from social networking sites (e.g., Myspace, Friendster, Facebook) to virtual world content (e.g., Second Life), from wikis (e.g., Wikipedia) to social-media blogs (e.g., BoingBoing, Engadget) and podcasting.

Web 2.0 changed, in the last few years, the vision of both personal and commercial websites, moving from large, closed and centralized repositories of static information to dynamic aggregators of heterogeneous contents, integrated into the Internet platform. This trend has been confirmed by the ever growing amount of API users can adopt to integrate their own applications and sites with the most important Web 2.0 applications, like, for example, YouTube or Flickr, implementing the so-called *architecture of participation*, where user interaction is encouraged in order to add value to the application itself.

Users can be effectively part of the development of Web 2.0 applications, by identifying the set of required features and validating the yet implemented ones, reducing the life cycle of applications and improving their usability, in a development approach known as *perpetual beta*.

Users interaction with Web 2.0 applications is exploited by Web services developers and providers because it also allows enriching the application contents by means of harnessing collective intelligence expressed by users. Tim O`Reilly (2007) shows how some of the most successful applications, which survived the transition between Web 1.0 and Web 2.0, are all characterized by a common property: the integration of users collective intelligence into their information flow. In particular the author presents the cases of Amazon, which obtained most of its success thanks to the books reviews written by users, and Google, whose ranking criteria, PageRank, is strongly based on the assumption that people used to link at most, in their personal websites, interesting and trusted documents.

The phenomenon of active participation has created a new platform for people to communicate with each other, to find new ways to build up and strengthen their own identity and to be a part of a group and participate to its evolution; it has been implemented by means of a new ease to use authoring tools, like the platforms for blogging (WordPress), social networking (MySpace, FaceBook) or media sharing (YouTube, Flickr). In addition to new graphical interfaces, Web 2.0 applications introduce the new concept of the *syndication*. Syndication is defined as a service used to notify to a set of subscribers the updates, which take place on a Web 2.0 content, such as the event generated by the publication of a new article into a personal blog. Syndication acts not only as a tool for resource monitoring but also

as a key element in achieving of the integration between heterogeneous data available on different sources. A typical example of this approach to information access is represented by blogging, one of the most common activities introduced by Web 2.0 philosophy. Syndication can enrich a blog by transforming it into a *live (or incremental) web site* (Skrenta, 2005) an entity able to interact with subscribers in order to notify them updates but also to act as a subscriber itself to integrate information coming from other syndicated sources.

Web 2.0 enables users to provide content as well as metadata, and to interact and sharing, producing, as side effect, information explosion and overload and highlighting some limitations, as lacking of accuracy of the retrieval tools and difficulty to a create adaptive filtering mechanisms with respect to user information needs and profile.

The idea of the Semantic Web (also called *Web 3.0*) is to apply semantic technologies in order to fill the knowledge gap between human and machine; it effectively moves from a feature-based representation of information (e.g. the keyword-based representation of textual contents or the level-histogram representation adopted to achieve retrieval of images and multimedia contents) to a knowledge representation, based on a common set of shared ontologies and reasoning rules. Different authors used the term Web 3.0 in order to represent the features related with application interoperability, ubiquitous and mobile computing, three-dimensional environments and semantic ontologies, indicating them as probable cornerstones of the Web of the next decade.

The research related to semantic aspects concerns many different research fields of Artificial Intelligence, like machine learning, natural language processing, database reasoning and knowledge representation. New and even more sophisticated methods for analyzing text and processing natural language will allow to develop automatic semantic tools which are capable of filtering information on the basis of the topic, identifying and extracting specific data, understanding the polarity of an opinion written by a user and organizing documents for similarity. With the employment of these and other techniques, like social network analysis, will pave path to the development of new knowledge management models and tools with a specific focus.

This chapter is organized as follow: after a first section dedicated to the classification of UGC applications, we discuss open issues and limitations in accessing, analyzing and extracting UGC and we present in a separate section a brief survey of those systems that integrate some of the features related with semantic representation and extraction from UGC. In this context, we propose our improvements in the area of information filtering, knowledge representation and sentiment analysis. Finally we focus our attention on economical implications of Web 2.0 and future trends. Conclusions end the chapter.

CLASSIFICATION OF UGC APPLICATIONS

UGC, also referred as User-Created Content (UCC) or Consumer Generated Media (CGM), allows every user to be linked as author, editor, customer and/or distributor of contents. Its increasing success has been estimated in (Horrigan, 2006): 35% of U.S. Internet users (about 48 million American adults) have provided at least one UGC during 2006.

UGC is defined in (Vickery and Wunsch-Vincent, 2007) as *"any kind of published content, result of a not professional activity with creative effort"*. UGCs include blogs, wikis, digital video, Internet broadcasting, mobile phone photography and photograph sharing.

A classification of UGCs, partly based on the schema introduced in (Blackshaw, 2005), is reported in Table 1 and discussed in succession.

1. *Blogs*. Rich, unaided first-person narratives across a host of topics; allowing user to

Table 1. Classification of UGC applications

Blogs, Message boards and forums	WordPress, Technorati, Blogger
Review/rating sites	Amazon, Tripadvisor, Epinions, Yelp, Ebay
Clubs or groups, Photo and Video sharing	Flickr, YouTube, GoogleVideo, DailyMotion, MetaCafè, PodZinger
Social networking	LinkedIn, MySpace, Friendster, Facebook, SecondLife
Collaborative authoring	Wikipedia, Google Docs, PBWiki, SlideShare
Social bookmarking and knowledge sharing	CiteULike, Connotea, CiteSeerX, Del.ici.ous, SharingPapers

enrich the published posts with UGC coming form heterogeneous sources available on the Web. Blogs are one of the most powerful UGC media; more specifically the support of the syndication mechanism allows blogs to share updates each other and to improve and speed up the indexing activity of search engines;

2. *Message boards and forums*. Evolution of a previously available Web 1.0 communication tool, the bulletin board, empowered by web access and interface. Such media are focused on specific topics (e.g. politics, lifestyle), products (e.g. cars, computers) or brands. With respect to the blogging platforms, which implements a one author to many reviewers communication policy, forum are based on a set of users acting as authors, delegating if necessary the review activity to a subset of administering users. Forum platforms allow users to map their reputation by means of, for example, number of submitted messages, number of received answers or time spent interacting with the platform. Sites like Google Groups and Yahoo Message Board provides access to a collection of several different specialized message boards.
3. *Review/rating sites*. Repositories of user reviews with respect to a set of products (e.g. movies, automobiles), people or services. Users can provide and share their own experiences or evaluate the goodness and usefulness of the previously published contents provided by other users
4. *Clubs or groups*. Highly focused and specialized sites, whose access is limited to a small amount of participants, where different UGC media can be integrated in order to exploit the specific topic of interest.
5. *Photo and Video sharing*. Applications that allow users to publish their own multimedia contents, and share such data each other. Users can interact with the published contents by means of voting, tagging or aggregation with their own contents; in this way users add value to the available data enriching them with some sort of collective intelligence (White, 2006), which can be useful, for example, in content retrieval and recommendation based on tags.
6. *Collaborative authoring*. Applications like Wikipedia allow users to participate in a collaborative way at the development of new multimedia contents. Services like Google Docs and SlideShare allow many different users worldwide to share, edit and store a set of documents simultaneously.
7. *Social bookmarking and knowledge sharing*. Web 2.0 applications can be used by users to create, organize and share more complex kind of UGC, like conceptual maps or taxonomies, built connecting each other available simpler contents.

ACCESSING, ANALYSING AND EXTRACTING UGC: OPEN ISSUES

Web 2.0 is subject to several severe limitations related, in particular, to retrieval and organization of UGCs:

1. *Information explosion overload*. 44% of U.S. Internet users are content creators (Horrigan, 2006) and the blogosphere is doubling in size every 200 days and 120.000 new blogs are being created each day (Sifry, 2007). Available information retrieval mechanisms are based on a feature representation approach. Such an approach does not provide a full understanding of the content meaning (e.g. keyword matching in textual documents does not require any kind of semantic evaluation of the meaning expressed by the body of a given document) and, especially in the UGC environment, does not look at contents in relation with the other available contents shared by users. A user trying to satisfy a specific information need can be easily overwhelmed by the amount of retrieved contents (Carlson, 2003).
2. *Multimedia information overload*. The increasing availability of software tools for the creation of multimedia contents allows users to communicate in a more sophisticated way by using rich media, which worsen the problem of information retrieval. As a matter of fact, rich contents like video and audio blog, podcasts, video lessons, online radio and Web TV and online repositories of multimedia contents (10 hours of new video are uploaded every minute on YouTube (Sarno, 2008)) should be dealt with on the basis of their real contents and not only by using traditional methods i.e. by text descriptions or tags indicated by the users.
3. *Complexity in analyzing and managing an open corpus of documents*. UGC generates an open corpus of documents (Micarelli et al., 2007): these documents do not share a common ontology and can constantly change and expand, increasing in such a way the complexity of knowledge management and retrieval. Furthermore, the online participation of people generates information in the form of comments and conversations without a specific structure and often characterized by an informal language. One single Web page, indexed by a search engine as a single document, can host hundreds of opinions; this increases the difficulty of analysis and extraction of knowledge.
4. *Difficulty in measuring information trust and quality*. With the growing number of producers of contents, also the need to obtain a measure of credibility of online information becomes ever more pressing to maintain the accuracy of search engines. For example, the information contained in a discussion is produced by different individuals, usually anonymous, difficult to identify and whose credibility is hard to measure (Anderson, 2007).
5. *Lacking of personalization*. A few search engines provide mechanisms that can adapt to the user actions and information needs. As the amount of information provided to the user becomes larger, unnecessary information can lead to difficulties in fulfilling his/her specific information need. Personalized systems are aimed to overcome this overload problem by building, managing, and representing information adapted for individual users (Gauch et.al., 2007). In spite of the fact that personalized systems improves the systems efficiency, effectiveness and usability, the existing techniques for adaptation and personalization of contents and navigation have proven their success in the case of finite corpus of documents. But the use of these techniques for open corpus is still to develop (Brusilovsky and Henze, 2007; Brusilovsky and Tasso, 2004).

6. *Flatness of folksonomies*. Social tagging is a flat mechanism, often ambiguous. People preferences over selection of tags may change as new trends keep evolving, this uncontrolled vocabulary used in folksonomies is creating a situation where effective classification and information management is hindered or slowed down. At present we are lacking in specific models to highlight and organize emergent folksonomies (Dattolo et al. 2008; Noruzi, 2006).

TECHNIQUES FOR ACCESSING, ANALYSING AND EXTRACTING UGC

This section proposes a brief look of the most promising lines of research in information retrieval, machine learning and data mining fields, aimed at improving the understanding of the specific semantic of UGCs.

Intelligent Scraping Systems

The first problem that needs to be solved in order to extract information and knowledge from the UGC is to obtain the raw data (conversations on forums, posts on blogs, comments, etc.) on to which carry out the analysis. The syndication mechanism can provide an important advice in order to solve the problem of access to raw data, by means of integration with standards used for notification of update taking place on a UGC, like RSS, Atom or other feed protocols). However only part of the content delivery platforms available on the Web implements the syndication approach. On the other hand, when contents delivery is achieved by a traditional server-push, client-pull mechanism, a different approach to scraping must be adopted.

A scraping system browses automatically a set of heterogeneous sources, identifies new pieces of information (e.g. a newly published post into a blog), filters out the sections of the selected web page which do not carry any relevant data (e.g. ads, navigation bars) and extracts the information contained in the page (e.g. date, title, author). Traditional scraping activity is achieved by means of textual analysis of the source code of each selected web page; more specifically analysis can be exploited by means of regular expressions or navigation of the page representation as a tree structure. Both approaches to scraping activity are based on a manually defined set of knowledge, used to navigate automatically and extract the right information. More sophisticated approaches, based on machine learning, are aimed at understanding automatically the structure of relevant data into a set of web pages retrieved from a specific source (Reis et al., 2004).

Collaborative Filtering Services

Social filtering, used in the past for content recommendation (Resnick and Varian, 1997) and electronic commerce (Schafer et.al., 1999), has gained an important role as core technology of many Web 2.0 applications, caused partly by the availability of large community of users, which participate in a Web 2.0 environment. Social search engines employ people's contributions to determine the importance of information, in contrast with the traditional approaches based on keyword occurrences and link analysis ranking. Several different Web 2.0 systems implement collaborative filtering mechanisms:

1. systems of questions and answers (e.g. Yahoo! Answer, MSN Live Qn, Amazon AskVille, Yedda, Answerbag);
2. systems of social bookmarking for the organization of links (e.g. del.icio.us, Furl, Simpy) or those specialised on specific domains (e.g. Citeulike, devoted to tagging and organization of scientific papers);
3. systems for specialized and personalized research, whereby the users put their experience at the service of a specific domain and

suggest lists of relevant sources (Hammond et.al., 2005).

The search takes place only within a list of trusted sources suggested by the users. Systems of social bookmarks represent an attempt to improving Web search and to solve the problem of information overload (Yanbe et.al., 2007) but currently have too limited sizes to gain a significant impact (Heymann et.al., 2008).

Sentiment Analysis and Opinion Mining

One of the promising research fields concerning semantic evaluation of Web 2.0 contents is related with the activity of identification and classification of the author's emotional and private issues (also referred as subjectivity) (Wilson et.al., 2004). Subjectivity can be seen as a rating indicator able to evaluate the amount of subjective information expressed by the text.

Many factors influence the subjectivity expressed by the author of a text, such as thoughts, experiences, motivation and interests and, mainly, positive and negative sentiments; all these elements constitute the so-called *private state* of a person (Wiebe et.al., 2001).

The subjectivity identification task oriented on sentiments, in terms of expressed polarity, is defined *sentiment (or opinion polarity) analysis* (Salvetti et.al., 2004). It may also be seen as a specialized way to perform Information Extraction, focusing on specific entities carrying the semantic of subjectivity expressed by the author of a UGC.

SA can be specialized in a several different task, aimed at:

1. assigning a subjectivity score to an input content, classifying it as objective or subjective, with respect to a set of previously evaluated subjectivity clues (Wiebe et.al., 2004).
2. evaluating the polarity of opinions expressed in the contents labeled as subjective during the previous task. Many researches exploited this problem (Casoto et. al., 2008a; Casoto et. al., 2008c; Pang, 2002; Turney, 2002; Liu, 2005; Gamon, 2005) proposing several supervised and unsupervised approaches. In particular, all these researches experimentally prove how such polarity classifiers, based on machine learning, may reach satisfactory results in terms of precision when applied to restricted domain and domain dependent corpus.
3. monitoring, by means of *sentiment timelines,* the trends in opinions related with specific entities, like users, places, concepts etc.. Sentiment timelines, at the same time, can be targeted to analyze the set of opinions expressed by a given user over time. Such as representation may be used to inference hypothesis about the private state of the user and enrich the profile describing the user, in addition to the knowledge that arises from monitoring user's interactions with the network and its contents.

7. Examples of applications implementing the SA process applied to UGC are OpinMind (http://www.opinmind.com) and Swotti (http://www.swotti.com). Opinmind is an opinion-driven search engine, based on a crawler, aimed at identifying and extracting opinions from textual contents available on the Web. The extracted opinions are evaluated with respect to specific polarity-bearing terms and classified as positive or negative. Users can enquire the system; relevant results are ranked and represented in a two columns table separating positive form negative opinions concerning the submitted query, Swotti is similar to Opinmind but focused on a smaller domain, related with merchandising; Swotti extracts

opinions from customers review sites, evaluates them by means of simple sentiment analysis heuristics, and aggregates the results with commercial data retrieved from several sources, such as the image collection provided by Google.

Cognitive Filtering

Cognitive filtering applied to UGC is one of the solutions adoptable in order to overcome the problem of **information overload** when accessing Web 2.0 contents. In particular cognitive filtering can be seen as a set of techniques aimed at identifying a subset of relevant items from a set of heterogeneous information sources, like, for example, a review site, a blog or a UGC repository. Cognitive filtering has been during the last ten years the leading research field of our artificial intelligence laboratory (Casoto et. al., 2008b); more specifically a specific set of instruments, the ifMONITOR (http://ifportal24.infofactory.it) tools, have been developed in order to cope with the requirements expressed by users interested in information access.

Based on a multi-agent architecture, ifMONITOR is devoted to cognitive filtering of textual contents, retrieved from several different sources available on the Web (sites, repositories of structured information) of from specific digital libraries or collection of contents. The crawlers which constitute the lower level of the ifMONITOR architecture, described in detail in (Asnicar, 1997), browse the available sources and extract, by means of an integrated intelligent scraping system, potentially relevant pieces of textual information, filtering out ads and navigational markup. Extracted data is matched against a set of manually or automatically defined cognitive profiles; document relevant with respect to at least one profile is tagged and delivered to the upper levels of the architecture. Document matching is achieved by means of the IFT algorithm (Minio and Tasso, 1996). IFT is able to represent both cognitive profiles and input data as semantic networks, constituted by cells, representing concepts, and edges, representing semantic relations occurring between concepts. Actually ifMONITOR supports the following concept representations: single terms, multi-terms and stems. IfMONITOR evaluates the similarity between the representation provided by the IFT algorithm and tags the input data accordingly with the results.

Relevant documents can be provided, by means of a service oriented approach, to several applications devoted to document management and delivery or publication. User can interact with the collection of relevant documents by means of such tools. One of the latest improvements applied to ifMONITOR concerns the ability to automatically extract tags from the semantic network representation of a given document. In this way we are allowed to perform two different kind of Web 2.0 activities: **automatic tagging** of textual UGCs harvested from an heterogeneous set of sources, based on the relevant concepts appearing into the content and, consequently, publication of the relevant retrieved items into existing Web 2.0 platforms. This last approach allow us to adopt ifMONITOR as an intelligent data aggregator, able to merge relevant contents and share them by means of common used Web 2.0 applications.

HARNESSING UGC: FUTURE TRENDS AND ECONOMIC IMPLICATIONS

The development of UGC is characterized by important economic implications, discussed in next subsections.

New Hardware and Software Requirements

The hardware producers could not ignore the emergent needs of people to share their thoughts and their knowledge online; the consumer market witnessed the introduction of new gadgets (e.g. phones, digital cameras, PDA) endowed with special features for the integration with the new user generated media.

As an example let's think of the peculiar features of some devices for uploading contents directly to online contents aggregators such as YouTube and Flickr. Furthermore a lot of new software systems allow the use of a person's mobile device to access his preferred social network or to publish in his personal blog.

New software houses are been setup leveraging on new software tools (e.g. like iWeb) which aim at simplifying the creation of contents (also multimedia like video and podcast) and their quicker distribution on the net.

New Tools for the Exploitation of UGC on the Traditional Media

New businesses are been set up and allow users to employ their own digital contents to create paper publications and distribute them completely bypassing the traditional distribution channels.

We are here referring to systems like MyPublisher or Lulu, which allow users to create and sell paper books starting from the digital contents of a person's blog. A different economic impact of UGC on traditional media derives from the new possibility of producing and selling digital contents avoiding completely the traditional distribution and promotion systems. For example, it is ever more frequent that a person derives a book from his blog, which is then sold directly as a pdf on a person to person way of business (e.g. the books 'save the pixel' by Ben Hunt's or 'Getting Real' by 37signals).

Also the traditional media are now employing UGC to increase their value. For example iReport of CNN has a community with over 80.000 users (iReporters) who can submit their articles and enjoy visibility on the CNN online channel. By publishing more than 1000 articles every month it has become an online newspaper completely written by the users.

New Ways of Advertising

The contents produced by the users are progressively becoming a real media. This has other economic implications arising from their use by the sector of the online advertisement. The business models in this area are several and diverse: users which include in their blogs sponsored links (Google AdSense, Feedburner etc.); content aggregators which include Ads between the contents uploaded by the users (YouTube Video Ads); systems which organize open contests for the creation of new advertisement campaigns and pay the winners (OpenAd, BlogBang); communities of bloggers who are paid to write articles to promote specific products or services under cover (PayPerPost).

New Means to Exploit the Information Produced by the Users

One of the most important features of the UGC is the possibility to access and analyze the spontaneous conversations of the users deriving new strategic knowledge of value to various areas of companies and organizations more in general.

Thanks to the new technologies, being developed for the retrieval, monitoring and semantic analysis of discussions in web forums, of articles and comments in blogs, of documents, podcast and videos uploaded by the web users, it is possible to derive strategic information for different business functions.

Business Intelligence

In this area the applications are very numerous. Ever more often the rumors are born on the net and only later reported by traditional media. Companies and non-profit organizations can increase the knowledge base for their strategic decisions by monitoring the Web in search of information on competitors, market changes, new technologies, violation of intellectual property (Kassel, 2001). Insiders can publish online secured information about the company, a new prototype, new ideas and business strategies. The rumors spread on the net are then aggregated on web sites and specialized communities (e.g. MacRumors, AppleInsider). By monitoring this information it is possible to get insights on new products, materials and financial strategies of the competitors, to identify new potential competitors, to monitor the updates in their prices, and to detect the transformations in the market just in time.

Marketing

New forms of promotions are been developed within the so-called **word of mouth marketing** (Womma, 2006; Gillin, 2007), which employ the users' conversations to diffuse specific messages and values of the company. In this area, specific ``listening instruments'' are fundamental which allow to:

- identify online information sources to monitor and analyze a company's credibility and influence on the net;
- continuously monitor the opinions and conversation identifying just in time new relevant information;
- filter discussions on the basis of their contents, accessing their relevance and classifying, for example, by topic;
- analyze the polarity of opinions (positive or negative) on the basis of specific parameters of the brand product or service, which is being analyzed (e.g. for a cell phone, it is possible to classify the opinions on the basis on quality of display, durability of the battery, etc.)

Extract Relevant Information

New systems of **information extraction** can be used to detect the citations of concurrent products, like names, prices, people names, geographic locations (Pudota et. al., 2008). The information can then be organized so as to offer an immediate glance on the most frequent concepts, the most cited products, the price range, etc..

Identify the opinion leaders and influencers of the community by analyzing the structure of a conversation and detecting the most active person in a specific forum on a specific topic. By assessing the polarity of each post it is possible to identify users particularly close to a certain brand and active in its promotion. Users like these, called influencers, can be of great value both if included in an online promotional activity of a new product and if involved in its development, for example to test new prototypes.

Identify the chains of dangerous information, that is, discussions that include misinformation or negative opinions or real defaming campaigns by unsatisfied users or competitors who pretend they are simple surfers and who spoil the brand. Companies cannot ignore situations of this type. In this case, it is important to be endowed with instruments, which highlight the presence of this type of activities to quickly respond and reduce the risk of a viral diffusion of misleading and dangerous information.

To measure the effects of a company's marketing actions, that is, to obtain from the analysis of UGC a clear and measurable indication of the positioning relative to the competitors and to trace down the modifications day by day as a consequence of specific online and offline promotional activities.

Product Development

One of the most innovative developments of Web 2.0, is the attempt to employ the user innovator (VonHippel, 1988). The user innovator is a user with new ideas for a development of a new product who is being included in the value chain of the company and of the product lifecycle (Wikstrom, 1996) in order to gain useful knowledge for the improvement of an existing product or for the engineering of a new one. The philosophy of managing ideas external to the company and with the potential of bringing innovation it is called *Open Innovation* (Chesbrough, 2005): it represents a new model of co-engineering of innovations. In this model the user (very often an online user) is being involved in all or some of the following phases:

1. Preliminary conception of the idea. It is the phase of idea development of a new product or identification and analysis of new trends and needs to satisfy. The Web 2.0, characterized by the active participation of the users, is very useful and allows to employ traditional instruments such as forums, blogs and wikies to manage a community of users close to a certain brand and involve them in different activities to generate new ideas to improve existing products or inventing new ones.
2. Design and engineering. Thanks to new technologies of rapid prototyping it is possible to involve the users in the evaluation of prototype of products. In this case the community becomes a focus group aimed at producing ideas and improvement insights. The producing company keeps improving the product until it gets the consensus by the community, which is often made up of thousands of customers.
3. Production. In some rare cases it is possible to delegate to the user also the production of the product. In this case, the examples are manly of digital products like photo (iStockphoto), video (Shutterstock Footage), graphics (Monster Templates), applications (iPhone Apps MarketPlace), promotional campaigns (Openad). There are also marketplaces where those who produce innovative ideas aimed at solving specific companies' problems (innocentive) can gain money or communities involved in the conception, engineering and collaborative realization of new products (the oscar project).
4. Testing. A newly released product can be distributed to a number of users in order to be tested. Several are the examples in the filed of software systems and web services where the users can freely try a service or a software application and share their impressions in a reserved community.
5. Promotion. The word of mouth marketing is becoming an important promotional instrument of a product or service. This central idea is to give the users the freedom to report their own personal experience with the use of a specific product/service so that their enthusiasm influences many others.

In the product development area it is also necessary to cite the development of the personalization system of the product itself. Today it is possible to configure in a personalized way a car, to assembly online a PC by choosing the various parts on the basis of personal needs (Dell) or build a musical compilation, which can then be voted by the other users and sold (iTunes iMix). Moreover, there are also several online systems which produce and sell a product completely engineered by a user such as mugs or T-shirts, or even more sophisticated products such as furniture and various gadget assembled from more simple parts (ponoko.com).

CONCLUSION

The explosive growth of user generated content as the prevailing form on the Web has raised several questions for the most effective approaches to

processes it; in fact, current technologies (exploited in Web 2.0) are not at all adequate to solve basic fundamental problems which are present in Web 2.0 even more than they were in Web 1.0: information explosion and overload, accuracy of retrieval tools, adaptive personalization, semantics of (textual) information.

Metadata are available in the form of tags, reviews, comments and recommendations, and could become invaluable in helping highly variable quality of content that end users are expecting.

The concept of quality in Web 2.0 has changed with respect to the decentralized and collaborative nature of the available contents. The absence of a centralized authority able to grant the quality of information and the ever-growing amount of available contents are leading to the idea of *good enough information*, having not been validated formally by an expert but accepted by a community of thousands of inexpert or practitioners users. This has given a new platform to researchers and developers to explore innovative ways for designing specific models to highlight and organize these emergent metadata.

This chapter has presented some open issues and indicated some emergent and innovative research lines to solve them by means of more sophisticated approaches, based on intelligent Web 3.0 techniques, moving beyond key-word matching and databases, towards deeper natural language understanding, machine learning, knowledge representation, and knowledge bases.

In adding, in order to make more complete the discussion, we have highlighting economic implications of Web 2.0 and their roles in next future.

REFERENCES

Anderson, P. (2007). What is Web 2.0? Ideas, technologies, and implications for education. *Bristol: JISC Technology and Standards Watch*. Retrieved on June 19, 2007, from http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf

Asnicar, F., & Tasso, C. (1997). ifweb: A prototype of user model-based intelligent agent for document filtering and navigation in the World Wide Web. *Workshop on Adaptive Systems and User Modeling on the World Wide Web at the 6th International Conference on User Modeling* (pp. 3-11), Chia Laguna, Sardinia, Italy.

Blackshaw, P. (2005). The pocket guide to consumer-generated media. Retrieved on March 7, 2007, from http://www.clickz.com/showPage.html?page=3515576

Brusilovsky, P., & Henze, N. (2007). Open corpus adaptive educational hypermedia. In P. Brusilovsky, A. Kobsa & W. Nejdl (Eds.), *The adaptive Web: Methods and strategies of Web personalization* (pp. 671-696). Berlin, Heidelberg: Springer.

Brusilovsky, P., & Tasso, C. (2004). Preface to special issue on user modeling for Web information retrieval. *User Modeling and User-Adapted Interaction*, *14*(2-3), 147–157. doi:10.1023/B:USER.0000029016.80122.dd

Carlson, C. N. (2003). Information overload, retrieval strategies, and Internet user empowerment. In L. Haddon (Ed.), *The good, the bad, and the irrelevant (COST 269)* (pp.169-173). Helsinki, Finland.

Casoto, P., Dattolo, A., Ferrara, F., Omero, P., Pudota, N., & Tasso, C. (2008). Generating and sharing personal information spaces. *Adaptive Hypermedia and Adaptive Web-Based Systems: Adaptation for the Social Web Workshop* (pp. 14-23). Hannover, Germany.

Casoto, P., Dattolo, A., Omero, P., Pudota, N., & Tasso, C. (2008). Sentiment classification for the Italian language. *Italian Research Conference on Digital Libraries*. Padua, Italy.

Casoto, P., Dattolo, A., & Tasso, C. (2008). (to be published). Sentiment classification for the Italian language: A case study on movie reviews. *Journal of Internet Technology*.

Chesbrough, H. (2005). *Open innovation: The new imperative for creating and profiting from technology*. Harvard Way, Boston: Harvard Business School Press.

Dattolo, A., Duca, S., Tomasi, F., & Vitali, F. (2008). *Towards disambiguating social tagging systems.*

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In A. F. Famili, J. N. Kok, J. M. Peña, A. Siebes & A. Feelders (Eds.), *Advances in intelligent data analysis VI* (pp.121-132). Berlin, Heidelberg: Springer.

Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). User profiles for personalized information access. In P. Brusilovsky, A. Kobsa & W. Nejdl (Eds.), *The adaptive Web* (pp. 54-89). Berlin, Heidelberg: Springer.

Gillin, P. (2007). *The new influencers: A marketer's guide to the new social media.* Sanger, CA: Quill Driver Books, Word Dancer Press.

Hammond, T., Hannay, T., Lund, B., & Scott, J. (2005). Social bookmarking tools: A general review. *D-Lib Magazine, 11*(4). Retrieved on March 10, 2008, from http://www.dlib.org/dlib/april05/hammond/04hammond.html

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve Web search? In *The International Conference on Web Search and Web Data Mining* (pp.195-206). New York: ACM.

Horrigan, R. J. (2006). *Pew/Internet-home broadband adoption 2006*. Retrieved from http://www.pewinternet.org/pdfs/PIP Broadband trends2006.pdf

Kassel, A. (2001). *Internet monitoring and clipping: Strategies for public relations, marketing, and competitive intelligence*. Retrieved from http://www.cyberalert.com

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the Web. In *14th International Conference on World Wide Web* (pp. 342-351). New York: ACM.

Micarelli, A., Gasparetti, F., Sciarrone, F., & Gauch, S. (2007). Personalized search on the World Wide Web. In P. Brusilovsky, A. Kobsa & W. Nejdl (Eds.), *The adaptive Web* (pp. 195-230). Berlin, Heidelberg: Springer.

Minio, M., & Tasso, C. (1996). User modeling for information filtering on INTERNET services: Exploiting an extended version of the UMT shell. In *Workshop on User Modeling for Information Filtering on the World Wide Web*, Kailia-Kuna, HI.

Noruzi, A. (2006). Folksonomies: (Un)controlled vocabulary? *Knowledge Organization*, *33*(4), 199–203.

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *International Journal of Digital Economics*, *65*, 17–37.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In J. Hajic & Y. Matsumoto (Eds.), *Conference on Empirical Methods in Natural Language Processing* (pp.79-86). Philadelphia: Association for Computational Linguistics.

Pudota, N., Casoto, P., Dattolo, A., Omero, P., & Tasso, C. (2008). Towards bridging the gap between personalization and information extraction. *4th Italian Research Conference on Digital Libraries*, Padua, Italy.

Reis, D. C., Golgher, P. B., Silva, A. S., & Laender, A. F. (2004). Automatic Web news extraction using tree edit distance. In S. I. Feldman, M. Uretsky, M. Najork & C. E. Wills (Eds.), *13th International Conference on World Wide Web* (pp. 502-511). New York: ACM.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56–58. doi:10.1145/245108.245121

Salvetti, F., Lewis, S., & Reichenbach, C. (2004). Impact of lexical filtering on overall opinion polarity identification. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. Stanford University: AAAI Press.

Sarno, J. (2008). On YouTube, more and more of everything. *Los Angeles Times*. Retrieved on March 10, 2008, from http://www.latimes.com/technology/la-ca-webscout2mar02,1,712991.story

Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender system in e-commerce. In *Electronic Commerce: Proceedings of the 1st ACM Conference on Electronic Commerce* (pp.158-166). New York: ACM.

Sifry, D. (2007). The state of the live Web. *Technorati Releases*. Retrieved on March 13, 2008, from http://technorati.com/weblog/2007/04/328.html

Skrenta, R. (2005). The incremental Web. *Topix Weblog*. Retrieved on March 23, 2008, from http://blog.topix.com/archives/000066.html

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *40th Annual Meeting of the Association for Computational Linguistics* (pp. 417-424). Morristown, NJ: ACL.

Vickery, G., & Wunsch-Vincent, S. (2007). *Participative Web and user-created content: Web 2.0 wikis and social networking*. Paris: Organization for Economic.

Von Hippel, E. (1988). *The sources of innovation*. New York: Oxford University Press.

White, B. (2007). The implications of Web 2.0 on Web information systems. In J. Filipe, J. Cordeiro & V. Pedrosa (Eds.), *Web information systems and technologies* (pp. 3-7). Berlin, Heidelberg: Springer

Wiebe, J., Wilson, T., & Bell, M. (2001). Identifying collocations for recognizing opinions. In *ACL/EACL Workshop on Collocation* (pp. 24-31). Toulouse, France: ACL.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics*, *30*(3), 277–308. doi:10.1162/0891201041850885

Wikstrm, S. (1996). Value creation by company-consumer interaction. *Journal of Marketing Management*, *12*(5), 359–374.

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *AAAI-04, 21st Conference of the American Association for Artificial Intelligence* (pp.761-769). San Jose, CA: AAAI Press/The MIT Press.

Womma. (2007). *101: An introduction to word of mouth marketing*. Retrieved on October 10, 2007, from http://www.womma.org/content/womma wom101.pdf

Yanbe, Y., Jatowt, A., Nakamura, S., & Tanaka, K. (2007). Towards improving Web search by utilizing social bookmarks. In L. Baresi, P. Fraternali & G. J. Houben (Eds.), *Web engineering* (pp.343-357). Berlin, Heidelberg: Springer.

ADDITIONAL READING

Auray, N. (2007). Folksonomy: the New Way to Serendipity. *International Journal of Digital Economics*, *65*, 67–88.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., & Su, Z. (2007). Optimizing web search using social annotations. In *16th international Conference on World Wide Web* (pp. 501–510). New York, USA: ACM.

Barbry, E. (2007). Web 2.0: Nothing Changes… but Everything is Different. *International Journal of Digital Economics*, *65*, 91–103.

Boiy, E., Hens, P., Deschacht, K., & Moens, M.-F. (2007). Automatic Sentiment Analysis in On-line Text. In L. Chan & B. Martens (Eds.), *ELPUB2007 Conference on Electronic Publishing* (pp. 349-360), Vienna, Austria.

Caschera, M. C., & D'Ulizia, A. (2007). Information extraction based on personalization and contextualization models for multimodal data. In *18th International Conference on Database and Expert Systems Application* (pp. 114-118). Washington, DC, USA: IEEE Computer Society.

Cheong, H. J., & Morrison, M. A. (2008). Consumers' Reliance on Product Information and Recommendations Found in UGC. *Journal of Interactive Advertising*, *8*(2).

Chevalier, M., Julien, C., Soulé-Dupuy, C., & Vallès-Parlangeau, N. (2007). Personalized Information Access Through Flexible and Interoperable Profiles. In M. Weske, M-S Hacid, & C. Godart (Eds.), *Web Information Systems Engineering – WISE 2007 Workshops* (pp. 374-385). Berlin, Heidelberg: Springer.

Chirita, P. A., Costache, S., Handschuh, S., & Nejdl, W. (2007). Ptag: Large scale automatic generation of personalized annotation tags for the web. In *17th international conference on World Wide Web* (pp.845-854). New York, USA: ACM.

Christiaens, S. (2006). Metadata Mechanisms: From Ontology to Folksonomy ... and Back.

Churcharoenkrung, N., Kim, Y. S., & Kang, B. H. B. H., (2005). Dynamic Web Content Filtering Based on User's Knowledge. In *International Conference on Information Technology* (pp. 184-188). Los Alamitos, CA, USA:IEEE Computer Society.

Damianos, L., Griffith, J., Cuomo, D., Hirst, D., & Smallwood, J. (2006, May). *Onomi: social bookmarking on a corporate intranet*. Paper presented at Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland.

Dattolo, A., Ferrara, F., & Tasso, C. (2009). Supporting Personalized User Concept Spaces and Recommendations for a Publication Sharing System. Geert-Jan Houben, Gord I. McCalla, Fabio Pianesi, Massimo Zancanaro (Eds.): *User Modeling, Adaptation, and Personalization, 17th International Conference, UMAP 2009*, formerly UM and AH, Trento, Italy, June 22-26, 2009. Proceedings. *Lecture Notes in Computer Science* (5535) Springer 2009, ISBN 978-3-642-02246-3, pp. 325-330.

Dattolo, A., Tasso, C., Farzan, R., Kleanthous, S., Bueno Vallejo, D., & Vassileva, J. (Eds.). (2009). *Proceedings of International Workshop on Adaptation and Personalization for Web 2.0* (AP- WEB 2.0 2009), Trento, Italy, June 22, 2009, CEUR Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-485.

Fabian Abel, F., Frank, M., Henze, N., Krause, D., Plappert, D., & Siehndel, P. (2007). GroupMe! - Where Semantic Web meets Web 2.0. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon et al. (Eds.), *The Semantic Web* (pp. 871-878). Berlin, Heidelberg: Springer.

Golder, S. & Huberman, B. A. (2005). The structure of collaborative tagging systems. *CoRR*, abs/cs/0508082.

Graham, R., Eoff, B., & Caverlee, J. (2008). Plurality: a context-aware personalized tagging system. In *17th international Conference on World Wide Web* (pp. 1165-1166). New York, USA: ACM.

Hayman, S. (2007, June). *Folksonomies and tagging: New developments in social bookmarking*. Paper presented at Ark Group Conference: Developing and Improving Classification Schemes, Rydges World Square, Sydney.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. doi:10.1145/963770.963772

In, R. Meersman, Z. Tari, P. Herrero (Eds.),*On the Move to Meaningful Internet Systems2006: OTM 2006 Workshops* (pp.199-207). Berlin, Heidelberg: Springer.

Krishnamurthy, S. (2008). Advertising with User-Generated Content: A Framework and Research Agenda. *Journal of Interactive Advertising*, *8*(2).

Lee, S., & Yong, H.-S. (2005). Web Personalization: My Own Web Based on Open Content Platform. In A.H.H. Ngu, M. Kitsuregawa, E. J. Neuhold, J-Y. Chung & Q. Z. Sheng (Eds.)

McDowell, L. K., & Cafarella, M. (2006). Ontology-Driven Information Extraction with OntoSyphon. In I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika et al (Eds.), *The Semantic Web - ISWC 2006* (pp. 428-444). Berlin, Heidelberg: Springer.

Mehta, B., Hofmann, T., & Nejdl, W. (2007). Robust collaborative filtering. In *ACM conference on Recommender systems* (pp. 49-56). New York, USA: ACM.

Mikroyannidis, A. (2007). Toward a Social Semantic Web. *Computer*, *40*(11), 113–115. doi:10.1109/MC.2007.405

Mizzaro, S., & Tasso, C. (2002). Personalization techniques in the TIPS Project: The Cognitive Filtering Module and the Information Retrieval Assistant. In S. Mizzaro & C. Tasso (eds.), *Personalization Techniques in Electronic Publishing on the Web: Trends and Perspectives - AH2002 Workshop* (pp. 89-93). Universidad de Màlaga, Spain.

Moschitti, A., Morarescu, P., & Harabagiu, S. M. (2003). Open Domain Information Extraction via Automatic Semantic Labeling. In I.Russell & S. M.Haller (Eds.), *Sixteenth International Florida Artificial Intelligence Research Society Conference* (pp. 397-401). Florida, USA: AAAI Press.

Noll, M. G., & Meinel, C. (2007).Web Search Personalization Via Social Bookmarking and Tagging.In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon et al. (Eds.), *The Semantic Web* (pp. 367-380). Berlin, Heidelberg: Springer.

Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Saggion, H., Funk, A., Maynard, D., & Bontcheva, K. (2007). Ontology-based Information Extraction for Business Intelligence. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon et al. (Eds.), *The Semantic Web* (pp. 843-856). Berlin, Heidelberg: Springer.

WebInformation Systems Engineering – WISE2005 (pp. 731-739). Berlin, Heidelberg: Springer.

Wu, X., Zhang, L., & Yu, Y. (2006). Exploring social annotations for the semantic web. In *15th International conference on World Wide Web* (pp. 417-426). New York, USA: ACM Press.

Xu, F., & Krieger, H. U. (2003). Integrating Shallow and Deep NLP for Information Extraction. In *Recent Advances In Natural Language Processing*, Borovets, Bulgaria.

Yuefeng, Li., & Ning, Z. (2004). Web Mining Model and Its Applications for Information Gathering. *Knowledge-Based Systems*, *17*(5-6), 207–217. doi:10.1016/j.knosys.2004.05.002

Zhu, T., Greiner, R., & Haubl, G. (2003). Learning a model of a web user's interests. In P. Brusilovsky, A. Corbett, & F. Rosis (Eds.), *User Modeling 2003* (pp. 65-75). Berlin, Heidelberg: Springer.

KEY TERMS AND DEFINITIONS

Business Intelligence: Broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions

Cognitive Filtering: Technique in which the description of a document is matched against a user profile where descriptions relate to static autonomous properties.

Collective Intelligence: Natural product of the independent opinions or behaviors of diverse individuals or groups in a decentralized system (flock, market, guessing game) that aggregates those opinions or behaviors. It is the intelligence of a collective, which arises from one or more sources

Folksonomies: Contraction of folk (person) and taxonomy, a folksonomy is a decentralized, social approach to creating classification data (metadata)

Information Extraction: The act of automatically extracting structured information, i.e. categorized and contextually and semantically well-defined data, from unstructured machine-readable documents

Ontology: An ontology is a collection of concepts and relations among them, based on the principles of classes, identified by categories, properties that are different aspects of the class and instances that are the things

Opinion Mining (Sentiment Mining, Opinion/Sentiment Extraction): Area of research that attempts to make automatic systems to determine human opinion from text written in natural language

Semantic Web: Abstract representation of data on the World Wide Web, based on the RDF standards. It is an extension of the current Web that provides an easier way to find, share, reuse and combine information more easily

User Generated Content (UGC): UGC refers to various kinds of media content, publicly available, that are produced by end-users. It reflects the expansion of media production through new technologies that are accessible and affordable to the general public these include digital video blogging, podcasting, news, gossip, research, mobile phone photography and wikis. In addition to these technologies, user generated content may also employ a combination of open source, free software, and flexible licensing or related agreements to further diminish the barriers to collaboration, skill-building and discovery