CHAPTER 2

**LITERATURE REVIEW**

The focus of this chapter is to review the prominent and relevant research that has been undertaken related to the proposed approach. Section 2.1 discusses the evolution of web giving an overview of the opportunities offered and the challenges associated with Web 1.0, Web 2.0 and Web 3.0 marking the clear distinction between them. This is followed by section 2.2 which elaborates on the various services and applications being offered by Web 2.0. The key focus of section 2.3 is to understand the concept of collaboration and participation. Section 2.4 discusses about the virtual community and how it differs from the other Web 2.0 communities. Section 2.5 provides a detail study of recommender systems which is followed by the study about the collaborative filtering technique in Section 2.6. Finally, Section 2.7 provides a summary of the literature review.

**2.1 Evolution of Web**

During the last decade, the World Wide Web has evolved into a large worldwide network as announced by many computer experts in the early 1990´s. Many people agree on Web evolution, but few people have thoughtfully studied its principles, i.e. why and how the Web evolves. Web evolution is supposed to be a major branch of Web Science.

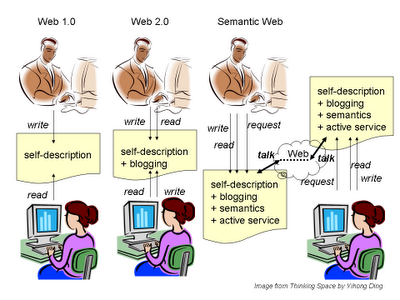
One of the early attempts of formalizing the concept of evolution on the Web was done by [Tim Berners-Lee](http://www.w3.org/People/Berners-Lee/) in 1989, the father of World Wide Web. In 1998, he explained the importance of [evolvability](http://www.w3.org/DesignIssues/Evolution.html) of Web technology. The first concept of Tim Berners-Lee that everyone can contribute was not fulfilled. Although millions of individuals were able to use the Internet, only a small percentage was capable to generate content. The main part of the online users was reading and consuming static pages. In other words, primarily technological aspects dominated the kind of internet access. From this point of view a World Wide Web network is only a connection from static internet pages prepared by some few web developers. The interaction or communication between the typical users was limited by the use of email, discussion forums and also chats. But currently something happens.

The last decade, with the constant inflation of the World Wide Web [1] and the familiarization of the users with the Internet, has generated all the necessary preconditions for a wide adaptation of the basic Internet as a generic exchange platform, where any user becomes a content provider i.e. the kind of web use dramatically changed. This came with the advent of Web 2.0[2], which is also known as Read/ Write Web. Web 2.0, first coined by Tim O´Reilly in 2004, helps the typical user to contribute and “The user is the content” is its most popular slogan. The popularity of Web 2.0 grows within all its applications. This new collaborative Web (called Web 2.0), extended by Web-based 'collaboration-ware' technologies like comments, blogs [3] and wikis [4], hosts successful sites like Myspace [5], Facebook [6] or Orkut [7], that allow to build social networks based on professional relationship, interests, etc. The term ‘Web 2.0’ is defined as the innovative use of the World Wide Web to expand social and business outreach and to exploit collective intelligence from the community. It advocates the Web architecture that promotes users’ participation and collaboration and acts as a basic platform for users to share, contribute, review and enhance information resources. Flickr [8] and YouTube [9] provide virtually unlimited media repositories for users to share photos and videos respectively. Collaboratively edited by any Web users, Wikipedia [4] has become one of the most resourceful encyclopedias in the world. The use of Weblogs, Wikis, Podcasts, and Social Bookmarking are summarized as Social Network.

**Figure 2.1:** Web-based 'collaboration-ware'

[](http://bp3.blogger.com/_U9YMKUF9sOg/Ru1smnFymAI/AAAAAAAAAGw/ECO1M0bg1Gg/s1600-h/web-evolution.png)

**Figure 2.2:** A Simple Picture Web Evolution

This picture above shows a simple abstraction of web evolution.

* **Web 1.0 – The World Wide Web**

The traditional World Wide Web, also known as Web 1.0, refers to the original information-oriented web. Web 1.0 is a Read-or-Write Web. In particular, authors of web pages write down what they want to share and then publish it online. Web readers can watch these web pages and subjectively comprehend the meanings. Unless writers willingly release their contact information in their authored web pages, the link between writers and readers is generally disconnected on Web 1.0. By leaving public contact information, however, writers have to disclose their private identities (such as emails, phone numbers, or mailing addresses). In short, Web 1.0 connects people to a public, shared environment --- World Wide Web. But Web 1.0 does not facilitate direct communication between web readers and writers. In other words, Web 1.0= Websites, E-mail newsletters and “Donate-now” buttons. It is one person or organization pushing content out to many people via websites and e-mail newsletters. It is a one-way communication and the donation process is not interactive or public. One donates and then receives a “Thank You” email.

* **Web 2.0 – The Social Web**

The second stage of web evolution is Web 2.0. The term itself was coined by Dale Dougherty in 2004 and popularized by [Tim O'Reilly](http://tim.oreilly.com/).It refers to the social web. It's a loose grouping of newer generation social technologies, whose users are actively involved in communicating and collaborating with each other as they build connections and communities across the web[10]. Web 2.0 is a Read/Write Web. At Web 2.0, not only writers but also readers can both read and write to a same [web space](http://yihongs-research.blogspot.com/2007/09/web-space.html). This advance allows establishing friendly social communication among web users without obligated disclosure of private identities. Hence it significantly increases the participating interest of web users. Normal web readers (not necessarily being a standard web author simultaneously) then have a handy way of telling their viewpoints without the need of disclosing who they are. The link between web readers and writers becomes generally connected, though many of the specific connections are still anonymous. Whether there is default direction communication between web readers and writers is a fundamental distinction between Web 1.0 and Web 2.0. In short, Web 2.0 not only connects individual users to the web, but also connects these individual users together. It fixes the previous disconnection between web readers and writers. In other words, Web 2.0 = Blogs, Wikis, Social networking sites. It is the beginning of two-way communication in the online public commons. People can post comments and converse with an organization in public for all to see. It’s one person or organization publishing content to many on social networking sites who then re-publish the content to their friends, fans, followers, connections, etc. We can also say that, here donation process is a public experience unlike in Web 1.0. Friends, fans, followers, connections, etc. on social networking sites see the giving and fundraising activity through widgets, apps, and peer-to-peer fundraising tools, like [fundraising pages](http://www.mercycorps.org/fundraising).

* **Web 3.0 - The Semantic Web**

The third stage of web evolution is **Web 3.0.** We don't know precisely what this stage of web evolution is at this moment. It refers to the currently evolving version of the web. There are different conceptions of Web 3.0. Some see Web 3.0 as the **semantic web (or the meaning of data)**, few others see it as a personalization (eg. iGoogle), and many of them consider it as an **intelligent web**, where software agents will collate and integrate information to give "intelligent" responses to human operators. This idea is associated with Tim Berners-Lee, the founder of the World Wide Web. Following the last two paradigms, an ideal semantic web is a Read/Write/Request Web. The fundamental change is still at web space. A web space will be no longer a simple web page as on Web 1.0. Neither will a web space still be a Web-2.0-style blog/wiki that facilitates only human communications. Every ideal semantic web space will become a little thinking space. It contains owner-approved machine-processable semantics. Based on these semantics, an ideal semantic web space can actively and proactively execute owner-specified requests by themselves and communicate with other semantic web spaces. By this augmentation, a semantic web space simultaneously is also a living machine agent. We had a name for this type of semantic web spaces as Active Semantic Space (ASpaces). In short, Semantic Web, when it is realized, will connect virtual representatives of real people who use the World Wide Web. It thus, will significantly facilitate the exploration of web resources.

A practical semantic web requires every web user to have a web space by himself . Though it looks abnormal at first glimpse, this requirement is indeed fundamental. It is impossible to imagine that humans still need to perform every request by themselves on a semantic web. Every semantic web space is a little agent. So every semantic web user must have a web space. The emergence of semantic web will eventually eliminate the distinction between readers and writers on the web. Every human web user must simultaneously be a reader, a writer, and a requester; or maybe we should rename them to be web participators. **In other words, Web 3.0 = Mobile Websites, Text Campaigns and Smartphone Apps.** Web 3.0 is all of the above except that the web experience is no longer limited to desktop and laptop computers while stationary in one place. It’s the Internet on the go fueled by mobile phones and tablets. [Mobile websites](http://nonprofitorgs.wordpress.com/2010/01/25/10-nonprofit-mobile-websites/) must be designed to be easily read on mobile devices. Group text campaigns function like e-mail newsletters in Web 1.0 to drive traffic to the user’s mobile website. Text-to-Give technology allows quick, easy donations on one’s mobile phone inspired by urgent calls to actions. Smartphone Apps enable content to be published and shared easily while on the go. Effectively donating via Smartphone Apps doesn’t exist yet, but its coming very soon.

In summary, Web 1.0 connects real people to the World Wide Web. Web 2.0 connects real people who use the World Wide Web. The future semantic web, however, will connect virtual representatives of real people who use the World Wide Web. This is a simple story of web evolution.

**Web 1.0 + Web 2.0 + Web 3.0 = Integrated Web Communications**

**Note:**

What’s important to understand is that all three eras of the web are complimentary and build and serve one another, rather than replace one another. They can also overlap. One uses Web 2.0 tools to drive traffic to the website, to build the e-mail newsletter list, and to increase visits to Donate Now buttons. Also, one uses his Web 2.0 communities to launch Web 3.0 campaigns and uses Web 3.0 tools to grow communities on social networking sites and to send supporters and donors to mobile versions of e-mail newsletter “Subscribe” and “Donate Now” pages. And while many nonprofit communicators are overwhelmed by all these new tools, it’s important to understand that there has been a [paradigm shift](http://www.taketheleap.com/define.html) in web communications. Some supporters and donors still prefer to be engaged by your nonprofit Web 1.0 style. Others think “[e-mail is for old people](http://chronicle.com/article/E-Mail-is-for-Old-People/4169)” and consistently get most of their content and inspiration from social networking sites. Web 3.0 will organize the masses in ways never seen before through [geolocation](http://en.wikipedia.org/wiki/Geolocation) , group texting and mobile websites, and much of it will be done via Facebook, Twitter, MySpace and Foursquare Smartphone Apps.

|  |  |  |
| --- | --- | --- |
| **Web 1.0** | **Web 2.0** | **Web 3.0** |
| “The mostly read only web” | “The wildly read-write web” | “The portable personal web” |
| 45 million global users(1996) | 1 billion+ global users(2006) | Focussed on individual |
| Focussed on companies | Focussed on communities | Lifestream |
| Home pages | Blogs | Consolidating dynamic content |
| Owning content | Sharing content | The semantic web |
| Britannica online | Wikipedia | Widgets,drag & drop mashups |
| HTML, portals | XML,RSS | User behaviour(“me-onomy”) |
| Web forms | Web applications | iGoogle, NetVibes |
| Directories(“taxanomy”) | Tagging(“folksonomy”) | User engagement |
| Netscape | Google | Advertisement |
| Pages Views | Cost per click |  |
| Advertising | Word of mouth |  |

**Table 2.1:** Differences between Web 1.0, Web 2.0 and Web 3.0

**Bottom Line:** There’s no “One Fits All” communication tool or tool set anymore. Age, class, race, gender and location play huge roles now in how people want to receive information and calls to action from nonprofits. The good news is that all of these tools are now affordable for nonprofits (even mobile marketing tools!). It’s just a matter of keeping up and finding the staff time – and the right person on staff – to master Web 1.0, Web 2.0, and Web 3.0. Those nonprofits that do it best will be the most successful in sharing their mission and programs, creating social change, and securing and maintaining new donors.

**2.2 Web 2.0 services**

The term Web 2.0 was coined on the evolution of web and various web-applications. The term itself was coined by Dale Dougherty in 2004 and popularized by [Tim O'Reilly](http://tim.oreilly.com/), the founder of World Wide Web. It is not a new concept of web, but a new way of using the web which came after Web 1.0. Although the term suggests a new version of the World Wide Web, it does not refer to an update to any technical specifications, but rather to cumulative changes in the ways software developers and end-users use the web.

Some popular Web 2.0 tools are podcasting, blogs, [RSS](http://en.wikipedia.org/wiki/RSS), social bookmarking, social networking sites, folksonomies etc. Blogs, wikis and RSS are often held up as an exemplary manifestations of Web 2.0. A reader of a blog or a wiki is provided with tools to add a comment or even, in the case of the wiki, to edit the content. This is what we call ‘*The Read/Write Web*’. At Web 2.0, not only writers but also readers can both read and write to a same [web space](http://yihongs-research.blogspot.com/2007/09/web-space.html) which allows for friendly social communication among web users [10]. What is important to recognize is that the emergence of the Web 2.0 is not a technological revolution, it is a social revolution [11]. This statement means that nowadays the usability of the technology gets simpler and simpler so that we are not forced to learn to use them in a technological way, but in a social way. It is a loose grouping of newer generation social technologies, whose users are actively involved in communicating and collaborating with each other as they build connections and communities across the web. Examples of various social networking sites are myspace.com, friendster.com, facebook.com, multiply.com, tagged.com, twitter.com, etc. Therefore, it is also referred to as ‘*The Social Web*’. It marks the progression from static web pages to dynamic, interactive ones. A Web 2.0 site gives its users the free choice to interact and collaborate with each other which leads to sharing of information and resources among them. It provides a number of services and applications that facilitate the features such as interactive information sharing, interoperability, and user-centered design. It is most commonly referred to as ‘*The participatory Web*’ that lets its users to actively participate and contribute i.e. there is complete user-involvement. Users can post their comment on news stories , can give their reviews on any information provided online , can upload their photos , can Share digital videos, etc. in contrast to the other websites where users were limited to the passive viewing of content that was created for them. Web 2.0 is ‘*The User-focused Web*’ whereinthe user needs are catered as they can freely participate, organize, read, write & play online. Web 2.0 draws together the capabilities of client-side and server-side software, content syndication and the use of network protocols. Standards-oriented web browsers may use plug-ins and software extensions to handle the content and the user interactions. Web 2.0 sites provide users with information storage, creation, and dissemination capabilities that were not possible in the environment now known as "Web 1.0".

**Figure 2.3**



**Figure 2.4**



**Figure 2.5**

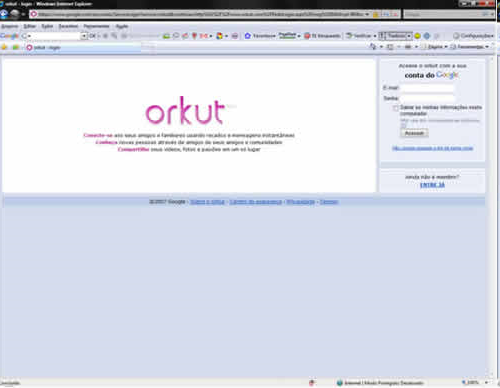


Various services of web 2.0 include social-networking sites, blogs, wikis, websites, podcasts**, vodcasts**, VoIP**,** RSS**,** folksonomiesand various other web applications.

**(a) Social-Networking** **Sites**

Social networking sites with [Facebook](http://www.facebook.com/) and [MySpace](http://www.myspace.com/) being the best-known - allow users to set up a personal profile page where they can post regular status updates, maintain links to contacts known as 'friends' through a variety of interactive channels, and assemble and display their interests in the form of texts, photos, videos, group memberships, and so on. This might involve drawing in other [Web 2.0](http://e-language.wikispaces.com/web2.0) tools like [RSS](http://e-language.wikispaces.com/rss) feeds, [folksonomies](http://e-language.wikispaces.com/folksonomies), photos, videos, etc, from [social sharing](http://e-language.wikispaces.com/social-sharing) sites. The field leader, Facebook, offers a huge range of optional third-party applications which can be used for entertainment, social, educational or professional purposes. It's also possible to set up group pages which can be accessed by 'members' or 'fans'. Social networking sites represent a fundamental shift from the content-oriented web (where webpages were usually about topics) to the person-oriented web (where webpages are about people).





**Figure 2.6 : ‘**Facebook’ and ‘Orkut’- The best-known social networking sites

**(b) Blogs**

Blogs [3] are like online journals where we can post updates - in the form of text, pictures, audio or video files. A blog can function as a reflective diary but it can also be the centerpiece of a conversation, since readers can leave comments for the blog's author and each other, forging connections and community around topics of mutual interest. The image at left shows the title bar.





**Figure 2.7 :** Blogging - The most recognized example of web 2.0

**(c) Wikis**

**Wikis** [4] are collaboratively authored websites, where anyone with a password can make alterations to unlocked sections. The image at left shows a section of the [homepage of this wiki](http://e-language.wikispaces.com/home) as it would appear to a user with editing rights. Wikis rely on the principle of collective intelligence and the notion that the product of collaborative work is often superior to what can be created by a single individual. Advantages for students include the ability to draft and redraft work collaboratively, with each contributor adding to and modifying the work of others. From an educational point of view, wikis are the perfect platform for social constructivist and community of practice approaches, and they are ideal for promoting a sense of a learning community. Feedback can be received from the entire internet (with a public wiki) or class peers (with a private wiki).

The main difference between a weblog and a wiki is that weblogs are personal whereas, wikis are mainly used for collaborative work. For example, if people work on the same documentation, a wiki system should be preferred.



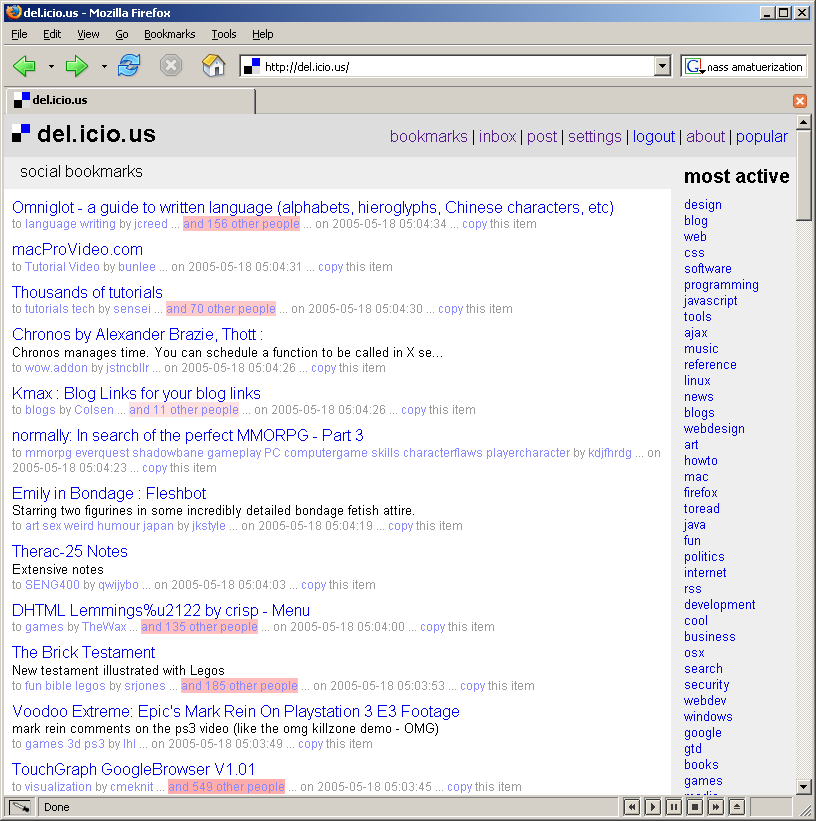
**Figure 2.8 :** Wikipedia- A collaborative dictionary being edited in real time by anyone

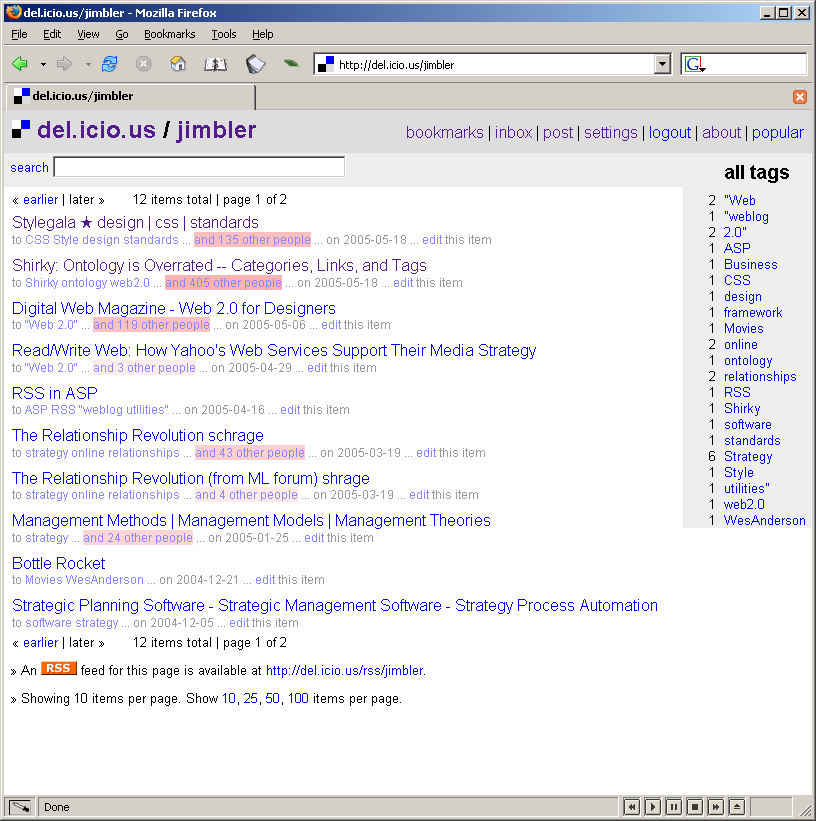
**(d) Podcasts**

Podcasts [11] are audio files, potentially with accompanying text and/or images - though if video is involved, they are referred to not as podcasts but as [vodcasts](http://e-language.wikispaces.com/vodcasting). They are distributed by syndication feeds such as [RSS](http://e-language.wikispaces.com/rss), with each new episode being downloaded to a computer to be played, or else transferred to a mobile device like an iPod or mp3 player. Once we have subscribed to a podcast, new episodes can be received automatically. They can be used, firstly, in a [Web 1.0](http://e-language.wikispaces.com/web1.0) manner, with teachers recording them and students simply being invited to listen. Such podcasts can range from lecture-style presentations to intensive language learning lessons. They offer many advantages in terms of recycling of material, whether that involves listening to a lecture a second or third time, or listening repeatedly to language learning materials. They have certain advantages over [vodcasts](http://e-language.wikispaces.com/vodcasting) as they offer the flexibility to engage in other activities while listening, unlike in vodcasts which require us to watch as well as listen. Podcasts can also be used, secondly, in a more [Web 2.0](http://e-language.wikispaces.com/web2.0) manner, with students being asked to create their own podcasts for publication to the web.

1. **Folksonomies**

Folksonomies [12], whose social nature is suggested by the inclusion of the word 'folk' in their name, are user-generated indexes of tagged sites found on the web, and typically underpin the process of social bookmarking. It is a spontaneous, collaborative work to categorize links by a community of users and users take control of organizing the content together. They represent a significant shift away from traditional top-down, hierarchical indexing systems for several reasons: the tagging process itself is organic rather than methodical or mechanical as we can simply add tags to relevant sites, there are no pre-set categories or subcategories and we can use whatever descriptive tags seem relevant and can even add new ones very often. This means that the index is flexible and extensible and the resulting index is usually presented in the non-linear form of a tag cloud. The larger a tag appears and the darker its font, the more often it has been used. Clicking on any tag will take us to a list of all of the sites which have been tagged with that particular term. Folksonomies depend on the [Web 2.0](http://e-language.wikispaces.com/web2.0) principle of collective intelligence, since they are a way of indexing distributed knowledge. The collaborative potential of folksonomies allows groups of people - whether educators, researchers or students - to work together to create folksonomies dependent on criteria negotiated and evaluated by members of the group.





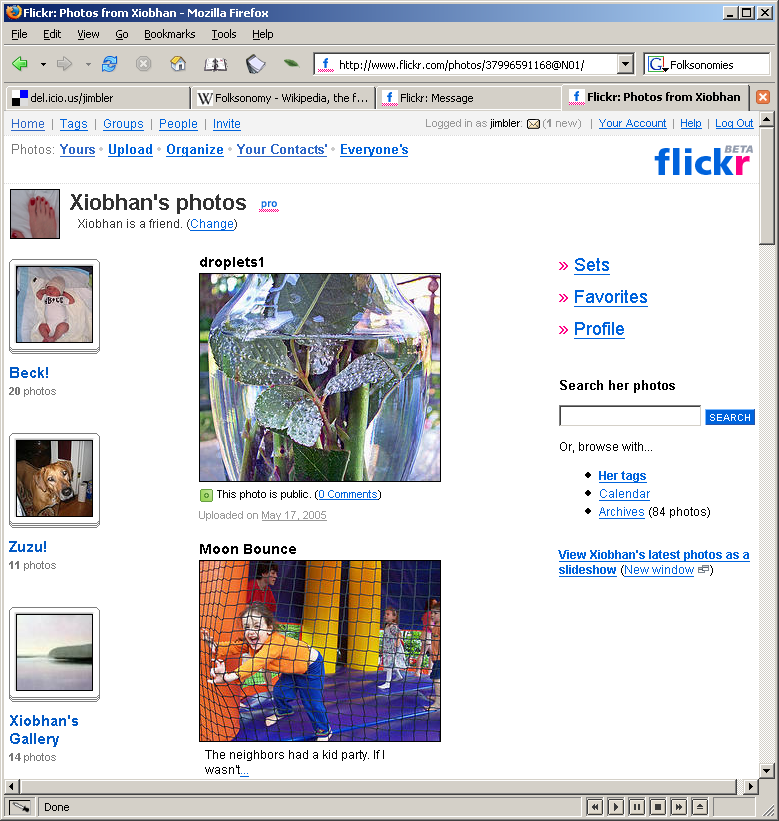
**Figure 2.9:** DEL.ICIO.US - An example of a site that uses a folksonomy to organize bookmarks

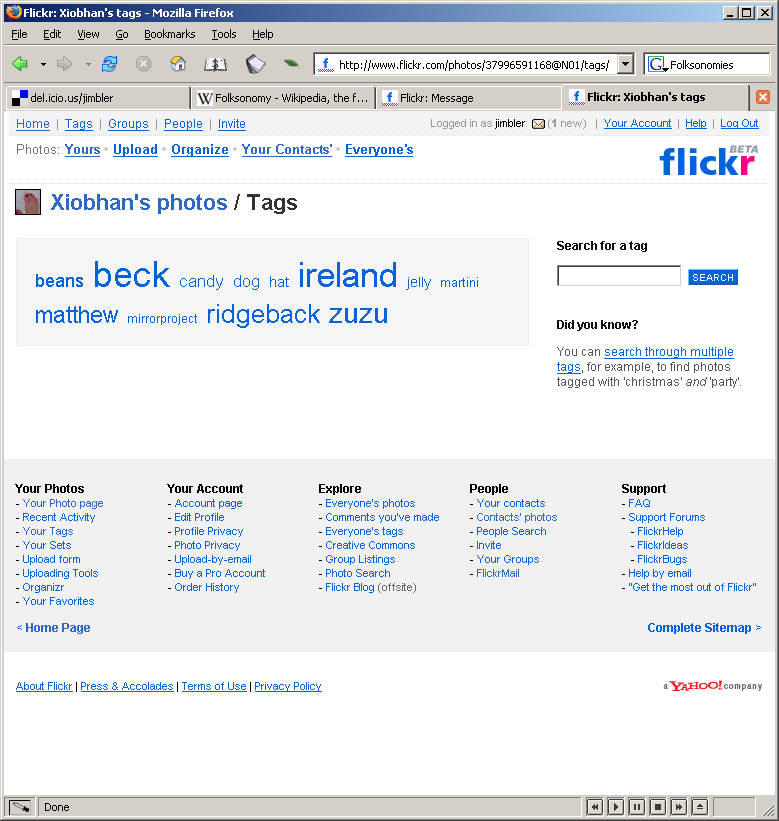
**(f) RSS**

RSS refers to Really Simple Syndication (RSS 2.0). They are often called as newsfeeds. Once RSS feeds are set up from the websites of interest, we receive automatic updates whenever those sites are updated. Like [folksonomies](http://e-language.wikispaces.com/folksonomies), RSS is about pulling together distributed content from across the web, but unlike folksonomies, RSS feeds provide a constant stream of up-to-date information in real-time from pre-selected sources. RSS feeds can be accessed either on the desktop of the computer or on the web itself and are an ideal way of keeping up-to-date information about subjects in which we are interested. Teachers can set up RSS feeds on relevant educational topics, or students can work together to set up group or class feeds on topics they are studying.

**(g) Websites**

Websites are made up of webpages (and often include a main page called a homepage). As vehicles for the delivery of information, websites, webpages or homepages have little to do with [Web 2.0](http://e-language.wikispaces.com/web2.0). In fact, static webpages are one of the most obvious features of [Web 1.0](http://e-language.wikispaces.com/web1.0). Over time, it has become easier and easier to create such pages. Nowadays, however, there is a whole new generation of websites, webpages and homepages which are dynamic rather than static and have a more [Web 2.0](http://e-language.wikispaces.com/web2.0) feel and orientation. These often draw in RSS feeds; draw in photos, videos, etc, from [social sharing](http://e-language.wikispaces.com/social-sharing) sites and allow user interaction through comments features, [discussion boards](http://e-language.wikispaces.com/asynchronous-discussion-boards) or [chat](http://e-language.wikispaces.com/synchronous-chat). For e.g. Flickr is one of the websites which combines a social network with user generated content. Users can work together to collaborate on photo projects and use each others’ tags to find new photos. Flickr also has an API for web services to integrate photo collections with blogs and other apps.





**Figure 2.10:** Flickr- A website which combines a social network with user generated content and used to collaborate on photo projects

**(h) VoIP**

**VoIP**, or Voice over Internet Protocol, refers to voice calls made over the internet. The best-known service provider is [Skype](http://www.skype.com/intl/en/helloagain.html), which also offers text and video channels.



**Figure 2.11 :** Skype **-** The best-known service provider

**(i) Vodcasts**

**Vodcasts** are much like [podcasts](http://e-language.wikispaces.com/podcasting) except that they involve video. Like podcasts, they are distributed on the web via syndicated feeds such as [RSS](http://e-language.wikispaces.com/rss), so once we have subscribed, new episodes can be downloaded automatically. Although they lack the flexibility of podcasts because users need to watch as well as listen, vodcasts have great potential for showcasing collaborative student projects. Although more professionally produced vodcasts may be created in media studios, or else vodcasts can also be made with any devices - including many mobile phones and digital cameras – which allows recording the video and uploading it to the net.

**2.3 Collaboration vs. Participation**

* + 1. **Collaboration**

Collaboration [13] refers to a recursive process where two or more people or organizations work together to realize shared goals. It is a deep and collective determination to reach an identical objective. It provides a basis for bringing together the knowledge, experience and skills of multiple team members to contribute to the development of a new product more effectively than individual team members performing their narrow tasks in support of product development.

This term came with the advent of Web 2.0 and holds a completely different meaning than participation which means sharing something in common with others. For example, participants can exchange videos, pictures and links, but the final purpose is not the achievement of a common objective. Collaboration is not a new term used in our work, and this term was introduced in the paper by Jin ,et.al. [14] where a framework was designed and developed for collecting, organizing and reusing a huge amount of information flooding on the Internet through group collaboration. The authors proposed a prototype system called Collaborative Information Browser, which provides functions such asreal-time collaborative information browsing, information organizing through adding comments, view-focused information retrieval and reusing and a platform for direct interaction among users. Also an object-oriented framework (SSVE) for collaborative group was designed by M. Linebarger, et.al*.* [15] which was highly suitable for a closely coupled small group collaboration in highly dynamic virtual environment. Numerous collaboration modes and mechanisms are provided by this framework along with various concurrency-control, interaction and interface modes. Further, a Web-based Framework for On-line Collaborative Learning was introduced by Shoul, et.al. [16] where the framework namely the Coco-Web (namely Collaborative Course-ware for the Web) was proposed to support design of course materials for on-line collaborative learning where course materials are classified into two categories: local documents and remote links. A local document can be either a HTML page, a piece of video, or a JPEG picture stored in the course-ware database, while a remote link points to some Web resources on the Internet.This framework utilizes the state-of-the-art technologies in semi-structured data management and personalization of web contents. Documents on different topics and at various levels could be organized in the framework, and structured to produce a personalized sequence of learning materials for each user (student). For on-line collaboration, the background and preferences of all users are described in their mathematical representations, and computed for appropriate grouping of collaborative users. The Internet resources could also be classified and recommended to the user based on the user's preferences and knowledge background. Also the paper by Lan and Huang [17] designs an efficient collaborative learning model to construct a consensus-based framework using the heterogeneous method and binary tree structure namely “Co-Tree” to organize the collaborative group. It uses the web-based CSCL approach, which is based on improving the collaborative learning by encouraging learners’ participation in learning activities which improves peer-interaction and leads to high learning satisfaction. Further, In the paper proposed by *Eryilmaz, et.al.* [18] , the collaborative knowledge is enhanced using two social technologies namely “Anchored discussion system” and “Non standard discussion system or stake holder controlled networking system” with the help of three methods as “multidimensional content analysis model”, “a sequential analysis of content” and “a social network analysis approach” considering few hypothesis . The final results of the paper suggested that the anchored discussion system seems to be particularly suited for collaborativeprocessing of research papers because it naturally directs students’ efforts towards theory oriented discussion of research papers, whereas the stakeholder system seems to be better suited to facilitate more individual oriented processing of research papers based on one’s prior knowledge as it helps individuals to organize their thoughts. In summary, the functional characteristics of the anchored discussion system place greater emphasis on collaborative knowledge construction within specific content. On the other hand, the stakeholder controlled networking system is oriented towards individual autonomy in knowledge construction while also mediating dialogue between peers to foster knowledge sharing.

Effective collaboration requires actions on multiple fronts:

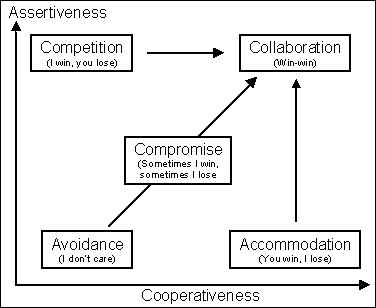
1. Resources
2. Teamwork
3. Process
4. Collaboration technology
5. **Resources**

Collaboration requires an early involvement of team members which requires that adequate resources must be available so that team members have time to effectively collaborate. For example; often there is a situation where a manufacturing engineer is assigned to a team, but is so busy managing resources that he or she does not have the time to effectively participate and collaborate. Thus, early involvement and the availability of resources is the basis to effectively collaborate and if collaboration is to work, management must provide adequate resources to support the project.

**ii. Teamwork**

Collaboration requires effective team work and team member cooperation. Team members must trust and respect one another and there must be open communication and a willingness to accept input from others. There should be defined team member responsibilities based on collaboration. Often, there are conflicting goals in product development and therefore decision-making must be based on a collaborative approach.

Different approaches to deal with an issue are mapped against the two axes i.e. Cooperativeness and Assertiveness as shown below in the figure:



**Figure 2.12 :** Different approaches to dealing with an issue

* A low degree of assertiveness and cooperativeness represents **avoidance** of an issue or the approach of "**I don’t care**".
* A high degree of cooperativeness and a low degree of assertion represents **accommodation** or the approach of "**You win, I lose**".
* A high degree of assertiveness and a low degree of cooperativeness represent **competition** or the approach of "**I win, you lose**".
* A moderate degree of both cooperativeness and assertiveness represents **compromise** or the approach of “**Sometimes I win and sometimes I lose** ".Many people believe that compromise is the most ideal approach to follow. But this is not so.
* A good team includes people that have strong beliefs i.e. a high degree of assertiveness and a high degree of cooperation. This represents the basis for **a collaborative** approach or the "**win-win**" approach. The key to the win-win approach is to creatively search for solutions that can mutually satisfy the needs of the team.

**iii. Process**

The process defines: what activities are performed, by who are they performed, when they are performed, and how they are performed. It presents the opportunity to enhance collaboration through parallel performance of activities which require early sharing of information and feedback, deliverables that require input, review and approval by other team members, early supplier involvement and less formal procedures and defined responsibilities (e.g., responsibility matrix) that require multiple team members involvement in activities and deliverables.

**iv. Collaboration Technology**

There are a variety of tools and technologies to facilitate communication and collaboration. Collaboration can happen synchronously where all participants view information and/or meet at the same time, or asynchronously where participants view information and provide feedback at different points in time. These tools and technologies include E-mail exchange of drawings, models and project information (asynchronous), teleconferencing and videoconferencing (synchronous), web-hosted meetings (synchronous), project hosting tools to create one pool of all released project documentation, with email alerts for updates (asynchronous), drawing viewing sites (intranet and web-based) with view and mark-up capabilities (asynchronous), CAD collaboration sessions (synchronous), workflow and groupware software (asynchronous), product data management, product information management, collaborative product commerce (generally asynchronous).

* + 1. **Participation**

Participation refers to sharing something in common with others. For example, participants can exchange videos, pictures and links, but the final purpose is not the achievement of a common objective. It is more like a “game” and it refers to different mechanisms for the public to express opinions - and ideally exert influence - regarding political, economic, management or other social decisions.

**2.4 Virtual Community**

Ongoing increase in wide-area network connectivity promise vastly augmented opportunities for collaboration and resource sharing. Now-a-days, various social networking sites like facebook, orkut, MySpace, Youtube have gained so much popularity and we cannot ignore them. They have become one of the most important application of Web 2.0 [19]. They allow people to build connection networks with other people in an easy and timely way and allow them to share various kinds of information and to use a set of services like picture sharing, blogs, wikis etc. Now , the most critical aspect is to use these social networks in a certain collaborative context. This gives rise to the “Virtual Community System” , most commonly referred as “On-line Communities”. The Internet now forms the basis for the constitution of online communities. Initially using the services of Web 2.0 we had the communities that were just following a participatory approach. These virtual communities offer a variety of functionalities, while they are proven to be efficient for educational and instructional goals [20].

A Virtual Community is a network of individuals who share a domain of interest about which they communicate online. The practitioners share resources (for example experiences, problems and solutions, tools, methodologies). According to the definition of Howard Rheingold in [21], virtual communities are social aggregations that emerge from the Net when enough people carry on public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace. Virtual community aims to meet the requirements of a virtual collaboration system with both autonomous and collaborative e-learning services, by supporting communities whose members interact and form groups based on their common interests. In other words, virtual community refers to groups of people who are in contact because they share same kind of knowledge and interests, corresponding with each other using interconnected computers in a cooperation process. It refers to the aggregation of people who have got similar interest and who work together to achieve a common goal or objective (collaboration). It adopts the specifications of Web 2.0 and provides three basic features viz. sharing of knowledge, transmission of knowledge and cooperation among its participants.

The paper by M. Brown, et.al.[22] proposes the different ways by which we can choose the right tools like wikis, blogs, chats etc. for the virtual team. The authors suggested a “needs-analysis” to be carried out containing the questions such as what do you need to accomplish, what are your current capabilities, which tool is appropriate for each task, and who is on your team. Also the use of wikis and blogs was highlighted in the paper by Kaldoudi, P. Bamidis, et.al.[23] wherein we find their use not just for the creation and promotion of information, but as active tools to support problem based learning (PBL) and active learning in the area of medicine. In this approach, students and instructors use the web as a virtual place to collaborate and create new knowledge and new educational experiences. Further, the paper by M. Zurada, et.al.[24] presented a framework and selected implementation issues of an Internet portal for the virtual community in computational intelligence and machine leaning (CIML). The main goals of the portal are to provide channels of communication between members of the community, to standardize ways in which resources are shared, and to build repository of these resources and to enable users of the portal easy access to various types of resources together with relevant software, references, datasets, methods, and their descriptions.

|  |  |  |
| --- | --- | --- |
| **WEB 2.0 COMMUNITIES** | **VIRTUAL COMMUNITIES** | |
| 1. Have generally no specific mission | | 1.) Have a specific mission-  People inside the VC work together to  achieve a common goal |
| 1. Interest of w2.0c is to aggregate as many persons as possible. | | 2.) VC objective is to aggregate just those  people interested in the community scope |
| 1. Have no tools specifically designed for e-learning activities, like  * SCORM Learning Object tools, * Video/webcast, agenda, * whiteboards, * Questionnaires & polls * exams, lists * mobile learning tools, * shared presentations | | 3.) Have various tools specifically designed for e-learning activities |
| 1. Have no inheritance mechanisms for communities-  * There is no hierarchy that ties different communities in one hierarchy. * Therefore ; no possibility of inheriting services from parents communities | | 4.) Have inheritance mechanisms for  communities-   * There is a hierarchy that ties different communities in one hierarchy and follows the idea of a “mesh” kind structure. |
| 5.) Have no sophisticated and specialized permissions on the services that they offer | | 1. Have sophisticated and specialized permissions on the services that they offer.  * The services are general applications activated by the manager of a VC that enable the users : * To communicate in synchronous and asynchronous way, * To publish contents, * To exchange files, * To coordinate events, etc |
| 6.)Have no inheritance mechanism for  permissions | | 1. Have inheritance mechanism for   permissions |
| 7.) Users have very simple roles & permission  schemas & cannot play different roles in the  same community. | | 1. Users of a VC can use the various services with different rights and duties  * In particular the rights and duties in the community are different from rights/duties for the services. |
| 8.) Structure of communities is flat;  Therefore no “propagation” tools can be  Implemented.(ex., send email to all member  of children communities) | | 1. VC follows the idea of a “mesh” kind structure.Services can take advantage of the “mesh” structure to provide some   interesting features like :   * “transversal wikis”, or “merged blogs” * One blog can be the “fusion” of all the blogs of children communities, or a wiki can take the definition transversally from all wikis in related communities. |

**Table 2.2 :** Differences between Web 2.0 communities (w2.0c) and Virtual Community

**2.5 Recommender System**

The advent and proliferation of the Internet and e-commerce has fueled the use of web for everyday tasks, right from sending e-mails, reading news articles, searching for information, viewing real-time stock quotes, and performing Internet banking to shopping online. Consequently, users spend most of their time in navigating through a large volume of information sources before identifying the nuggets of information they seek. Furthermore, as the web is growing complex on a continuous basis, the information overload problems faced by the users are also increasing at a rapid pace. Recommender systems [25] help in addressing this information overload problem by retrieving the information desired by the user based on his/her similar users' tastes and preferences.

Recommender systems are systems that provide recommendations to customers based on their past purchases, tastes, and preferences [26]. For instance, Amazon.com (www.amazon.com) site's features such as "Customers who bought" and "Book Matcher," and CDNow (www.amazon.com) site's features like "My CDNow" and "Album Advisor" are some of the typical examples of recommender systems. The "Customers who bought" feature of Amazon.com can be found on the information page for every book in their site. The basic principle of the "Customers who bought" feature is that: the recommendations are offered based on the books frequently purchased by customers who purchased the selected book.

Recommendations offered by recommender systems can be as simple as offering a web page (based on average ratings of web pages) to as complex as providing products in online shopping (by analyzing a customer's complex click and purchase histories).

**2.5.1 Common Technologies**

Text classification has been one of the key tools to automatically handle and organize text information for decades. In recent years, with more and more subjective information appearing on the internet , sentiment classification as a special case of text classification for subjective text is becoming a hotspot in many research fields including Natural Language Processing (NLP), Data Mining (DM), and Information Retrieval (IR). It is a complex process which requires more than just text mining techniques . It has been found that text mining algorithms on sentiment classification do not perform as well as that on traditional topic-based categorization as topics can be identified by keywords but sentiment would be expressed in a more subtle manner.

Unlike traditional topic-based text classification, sentiment classification concerns on the sentiment words and classifies the reviews into sentiment categories of positive or negative (or sometimes neutral). Supervised machine learning methods and unsupervised techniques have been applied for the task of sentiment classification.

Qiang Ye , et.al. [31] apply three different machine learning techniques (Naïve Bayes, Support Vector Machine and Character based N-gram model) on travel blogs . Experimental results show that standard machine learning techniques perform well on sentiment classification. Moreover, from the empirical findings and comparison result of the three machine learning methods, Support Vector Machine (SVM) achieves the best performance, while Naïve Bayes tends to be the worst one. ZHU Jian, et.al. [32] also investigate similar task with an individual model (i-model) based on Artificial Neural Network (ANN) to classify movie reviews into positive and negative using sentiment features, feature weight and prior knowledge base. A conclusion in their experiment is that that the accuracy of i-model is higher than that of SVM and Hidden Markov Model (HMM). Other related work of using supervised machine learning methods for sentiment classification can be found in the paper by Rui Xia, et.al. [33] wherein two types of feature sets ( part-of-speech based feature sets and word-relation based feature sets) , three well-known text classification algorithms ( Naïve Bayes, maximum entropy and SVM) and three types of ensembling methods (fixed combination, weighted combination and meta-classifier combination) are used to conduct a range of comparative experiments on five widely-used datasets. Minqing Hu and Bing Liu [34] propose an unsupervised approach to mine opinion features and determine the polarity of opinion sentences. Their work is performed in three steps:

(1) mining the product features and opinions that have been commented on by customers. First, they used association rule mining (ARM) to find all frequent item sets, after that they did feature pruning or compactness pruning and at last they extracted the nearby adjective words as opinion words at last; (2) identifying the opinion sentence in each review and deciding whether each opinion sentence is positive or negative, they used WordNet to predict the semantic orientations of opinion words; (3) summarizing the results. Based on the above paper, Qingliang Miao , et.al. [35] proposed a novel ranking mechanism which takes into account the temporal dimension of the opinion (temporal opinion quality, TOQ). Displaying the trend movement of sentiments along the time axis can be very useful in many applications as the opinions of people on a particular product or a topic may change over time. Other related work of using unsupervised machine learning technique for sentiment classification can be found in the paper by Yao Lu, et.al. [36] which takes the strength of user sentiments into consideration , which is important in measuring the overall quality of products and services. Sentiment strength of user reviews is estimated according to the strength of adverbs and adjectives expressed by users in their opinion phrases. The experimental results obtained are close to, and sometimes better than those of supervised classifiers. One of the key issues in sentiment classification is that the sentiments of text can be classified at varying levels of granularity viz. document level, sentence level, phrase level, word level [37]. The ability to classify sentiments on multiple levels is important since different applications have different needs. For example, a summarization system for product reviews might require polarity classification at the sentence or phrase level, a question answering system would most likely require the sentiment of paragraphs, and a system that determines which articles from an online news source are editorial in nature would require a document level analysis. Also, the classification decision at one level in the text influences the decision at another level in the text. The related work is found in the paper by Yi Hu, et.al. [38] which is based on sentiment classification at document level by modeling description of topical terms. It is based on the concept of constructing two sentence-level sentiment description models , namely positive and negative Topical Term Description Models(TTDM) , used for the sentiment classification at the sentence-level which in turn can be used to decide the whole document classification collectively. This way global document classification will benefit greatly as it has been established that sentence level classification improves document level analysis.Similarly, the paper by ZHAO Yan Yan, et.al.[39] improves the sentence sentiment classification by integrating Intra and Inter- document evidences. The authors suggested that determining the sentiment orientation of a review sentence requires more than just focusing on the features inside the sentence, especially for the sentences with ambiguous features inside a sentence. Features outside a sentence need to be explored, so as to interact with its inside feature to enhance the performance. Apart from the works of binary sentiment classification, there are also some works performed in terms of assigning the polarity strength to the adjectives. Williams and Anand [40] propose a knowledge-based approach to assign each adjective with polarity strength by measuring semantic distance between words of known polarity (seed words) and the target word according to WordNet [Appendix C], which is the most similar work to ours.

Some of the other popularly used techniques for sentiment classification :-

1.) One of the most commonly used techniques for recommendation generation is **collaborative filtering**. Collaborative filtering identifies a subset of users that have similar tastes and preferences to that of the target user to generate recommendations. More specifically, the collaborative filtering process involves three-stages: (1) computing similarity between the target user and all other users; (2) selecting a subset of collaborative users based on the similarity coefficients computed in Step 1; and (3) offering recommendations based on products liked by collaborative users.

2.) **Data mining** is another technique used for recommendation generation. Data mining is defined as a non-trivial process of extracting potentially useful, interesting, and actionable information from massive databases. Specific data mining techniques used for recommendation generation include association rule mining, clustering, web mining, or a combination of them.

3.) **Information retrieval** is yet another method for recommendation generation. There are a variety of shopping assistants available on the web (DealTime—www.dealtime.com; Shopping.com—www.shopping.com) that use information retrieval-based methods. These shopping assistants provide agent-based shopping support for customers. They take price and a set of product features as inputs, and match them with available products on the Internet to select a set of products of interest to the customer. These agents also provide services such as product ratings, customer reviews, price comparisons and details of product availability across stores. However, selecting suitable products in the vast Internet is a challenging problem. Other shopping agents available on the web such as Active Buyers' Guide (www.activebuyers guide.com) take into account the importance of product features in addition to the feature itself to select products of interest to the customers. In essence the shopping assistants or agents available on the web use a set of customer-desired features, and match the same with the available products on the web to select a set of products for recommendations. The recommendations generated in such systems are generally product variants rather than cross-category products as in collaborative filtering-based methods.

E-commerce sites use recommender systems to offer products of interest to the customer with a view to: (1) convert their browsers into buyers; and (2) identify cross-sell/up-sell opportunities for their buyers. Essentially, recommender systems help e-commerce sites earn more revenue for their site. Recommender systems attenuate the information overload problem on the web by offering products of interest to the customers. Currently, recommender systems are being used widely in the fields of movies, music, web pages, retailing, and so on.

In the preceding section, a general introduction to recommender systems was provided. Now, we will look at the design aspects of recommender systems. Various inputs, processes, and outputs needed for recommender system design are elaborated in detail. Subsequently, a discussion on recommender system types and recommender system services is presented.

**2.5.2 Recommender System Process**

Recommendation generation typically involves converting a set of inputs into recommendations using a pre-defined process [25]. A generic input-process-output model for personalized recommender systems is given in Figure 2.13. It provides a comprehensive set of inputs, processes, and outputs that are generally used in recommender systems. Different recommender systems have been developed in the past using different combinations of these inputs, processes, and outputs under various domains.

**2.5.2.1 Inputs for Recommender System**

Figure 2.13 gives a large set of inputs that can be used by recommender systems. However, the types of inputs for recommender systems can be broadly classified as explicit, implicit, and product information. Table 2.3 gives a summary of the types of input data used along with a few illustrative examples.

**DOMAIN**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **INPUT** |  | **PROCESS** |  | **OUTPUT** |
| 1.)Explicit Rating |  | 1.)Collaborative Filtering |  | 1.)Top-N Products |
| 2.)Customer Feedback |  |  |  |  |
| 3.)Demographic Profile |  | 2.)Data Mining |  | 2.)Average Rating |
| 4.)Ephemeral-Needs 5.)Purchase History |  | 3.)Information Retrieval |  | 3.)Most Popular |
| 6.)Navigation-History  7.)Product Taxonomy |  | 4.)Mathematical Models |  | 4.)Most Frequent |
| 8.)Product Attributes |  | 5.)Hybrid |  | 5.) Latest products |
| 9.)Product Descriptions |  |  |  |  |

**Figure 2.13 :** Generic Input—Process—Output Model of Recommender Systems

|  |  |  |
| --- | --- | --- |
| S.No. | Type of Data | Examples |
| 1 | Explicit data | Customer ratings  Feedback  Demographics  Psychographics  Ephemeral needs |
| 2 | Implicit data | Purchase history  Click history |
| 3 | Product information | Product taxonomy  Product attributes  Product descriptions |

**Table 2.3:** Types of Input Data Used in Recommender Systems

Explicit ratings are subjective ratings given by users of the system based on their tastes and preferences. For example, a customer who had purchased a book called "Transformative Organization across the Globe" gives a rating for his liking of the book as 4.5 on 5. These ratings are then used for providing recommendations to other users who have similar tastes and preferences. The types of explicit inputs used in recommender systems include explicit ratings; customer feedback, demographics, and ephemeral needs.

Implicit inputs or ratings are captured by the system automatically without any effort on the part of the users of the system. Customers' implicit preferences are captured using their purchase and navigation histories.

Product information includes details about the product such as product taxonomy, attributes, and descriptions. Sample product taxonomy for the product category of personal care and grooming is provided in Figure 2.14.Product attributes are the attributes of the product that can be found in the product database. For instance, Table 2.4 gives a list of product variants of color TVs. The table provides three features (make, size, and warranty) on eight attributes (LG, Samsung, Sony, 14", 21", 1 yr, 2 yr, 3 yr) for color TVs. Product P1 in the table means that it is of LG make and 14" in size and the warranty offered is 1 year. If the customer-specified attributes (say, LG make and 21" size) match the products in the database, then that set of products is offered as recommendations.

**Figure 2.14:** Product Taxonomy for personal care and grooming category

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Product ID** | **Make (F1)** | | | **Size (F2)** | | **Warranty (F3)** | | |
| **A1** | **A2** | **A3** | **A4** | **A5** | **A6** | **A7** | **A8** |
| **P1** | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| **P2** | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| **P3** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| **P4** | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| **P5** | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| **P6** | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| **…** |  |  |  |  |  |  |  |  |
| **…** |  |  |  |  |  |  |  |  |
| **P10** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |

**Table 2.4:** Product Database for Color TVs

**2.5.2.2 Recommendation Generation Process**

The processes adopted for generating recommendations (refer to Figure 2.13) include collaborative filtering (CF), data mining (DM), information retrieval (IR), and mathematical models.

1. Collaborative filtering (also referred to as social information filtering) [27] is one of the most widely applied techniques among them. The primary idea of the technique is to find a subset of users who have similar tastes and preferences to the target user and using this subset for offering recommendations. That is, given a target user, u, compute his n (say) similar users [u1, u2, …, un] and predict u’s preference based on the preferences of [u1, u2, …, un] . Collaborative filtering technique has been widely applied in many commercial systems. For example, online retailer Amazon.com uses collaborative filtering technique for recommendations such as "Customers who bought Book A, also purchased Book D, Book F, and Book L".
2. Data mining techniques used in recommender systems include association rule mining (ARM) and a combination of clustering, web mining (WM), and/or ARM.

ARM primarily identifies the set of products purchased together by the customers. For example, 80% of the customers who purchased Product X also purchased Product Y (the strength of the rule being 90%) is a typical association rule. Table 2.5 and Figure 2.15 provide customer purchase database and sample association rules. In Figure 2.15, each rule is specified with two parameters: support and confidence. Support is the frequency of occurrence of the items in the antecedent and consequent of the rule and confidence is the conditional probability that whenever the item in consequent is purchased, the item in antecedent is also purchased. In the recommender system context, association rule mining is performed on the purchase/click history of customers to offer as recommendation the related products purchased by customers.

|  |  |
| --- | --- |
| **Transaction ID** | **Products Purchased** |
| 1 | Beer, Milk, Bread, Cornflakes |
| 2 | Bread, Butter, Cornflakes |
| 3 | Bread, Butter, Beer, Diapers |
| 4 | Beer, Diapers |
| 5 | Milk, Cornflakes, Bread, Butter |
| 6 | Bread, Milk, Cornflakes |

**Table 2.5:** Customer Product Purchase Database

Beer => Diapers (33%, 67%)

Bread, Butter => Cornflakes (33%, 67%)

Milk => Cornflakes (50%, 100%)

**Figure 2.15:** Sample Association Rules Mined from Table 2.5

1. Information retrieval is yet another process used for offering recommendations. Here, the content information is mined using text retrieval methods to provide useful recommendations. Information retrieval-based methods also involve taking inputs from the customer in the form of attributes of the products. The customer-specified attributes are then matched with the products in the database to offer products of interest to the customer. For example, let us say a customer wishing to purchase a color TV specifies her requirements as Samsung make and 21" size; these requirements are matched with the products in the database (as in Table 2.4) and a set of products is offered as recommendations.
2. Various mathematical models used in the recommender systems include naive Bayes, Bayesian belief networks, dependency networks, and Markov processes .

**2.5.2.3 Outputs for Recommender System**

The outputs generated by the recommender system can be Top-N products (based on ranking of the products), average ratings (taking the average of the ratings provided by other customers), and the most popular, most frequent, or latest products. The choice of the output method depends on the particular needs of the application. A combination of outputs can also be generated by the recommender system.

The type of inputs and processes selected for a recommender system (refer Figure 2.13) largely depends on the domain under study. For instance, in a movie recommender system the user's explicit ratings for movies are taken for offering future recommendations. However, in a retailing situation where there are millions of users and products, the use of explicit ratings is not quite feasible.

**2.5.3 Types of Recommender System**

The types of recommender system services can be broadly classified into two categories: (1) personalized service and (2) non-personalized service.

In a personalized service, the users are offered customized recommendations based on their stated (explicit) or implied (implicit) needs. MyCDNow (www.amazon.com) and Match Maker are some of the examples of personalized services.

In a non-personalized recommendation service, the same sets of recommendations are offered irrespective of the user who requests the services. Non-personalized recommendations are generated based on what other users have said about the products on an average. For example, average ratings of buyers and sellers (eBay) and average customer ratings (Amazon.com) are some of the non-personalized recommendation services.

A lot of recommendation services (personalized/non-personalized) can be offered to users of the web. Here we discuss some of the plausible services in the context of online shopping. The purchase behavior of a customer in the online context can be well understood with the help of online shopping cycle shown in Figure 2.16. In a typical online shopping, a customer goes through a series of phases, viz., browsing through the products, selecting the products of their choice, placing the selected products in the shopping cart, and finally making a decision to purchase or not . We can identify three major phases in the shopping cycle: click-through phase (i.e., from the moment a customer visits the site to click on certain items to view more information about them), basket placement phase (from clicking on certain items to viewing its information to placing the items on the basket or shopping cart), and basket-to-purchase phase (i.e., from the moment at least one item is placed in the basket till the customer actually leaves or checks out of the site).

Browse through Products

Place products in the shopping Cart

Login to the store

Purchase the products in the cart

Click-through Phase

Basked-to- Purchase Phase

Basket Placement Phase

**Figure 2.16:** Examples of Book and Movie Recommender Systems

Recommender systems can be tailored to offer different services by developing an understanding of different phases of online shopping cycle. Some of the recommendation services that can be offered in each of the phases of online shopping cycle are discussed in the next section.

1. **Recommendation Services for Click-through Phase**

In the click-through phase, the user or customer has just visited the site and has not yet clicked on any items on the site to view its information. Hence, recommendations can be offered based on:

* User's past purchases or clicks (a personalized service); for example, a service like Gold Box Offers (www.amazon.com) could be provided.
* Keywords provided by the user in the search facility during his/her visit to the site (a non-personalized service).
* Wish lists or shopping lists, i.e., the products a customer buys regularly can be stored on a wish list and retrieved by him/her as and when required. This service would save a lot of time for the customer in searching the products they buy on a regular basis. So, whenever a customer visits the site, they can just add the items in their wish list to the shopping cart and add any additional products to the cart (a personalized service).
* Personalized suggestions of products (or categories of products) can also be offered based on the user's wish list.

1. **Recommendation Services for Basket Placement Phase**

In the basket placement phase, the user has clicked at least one item and viewed its detailed information. In this phase, the user navigates through the hierarchy of product (sub-) categories before they identify the product that suits their needs. So, the recommendations should be tailored toward assisting the customer in quickly identifying the products that, suit their tastes and preferences. Some of the services that can be offered in this phase include the following:

* E-store navigation, i.e., customer store navigation details, are used for tailoring the recommendations. Here the recommendations can be personalized or non-personalized based on whether the customer has sufficient click-throughs or not.
* Product assortments, i.e., the ordering of products when the customer views the information, can be varied based on his/ her individual profile. For instance, customer's favorite products (or categories) could be placed on top of the list (a personalized service).

1. **Recommendation Services for Basket-to-purchase Phase**

In this phase, a customer has placed at least one product in the basket. The customer can do a set of actions in this situation. That is, he can: (1) leave the site abruptly; (2) navigate through the site for selecting more products; and (3) proceed to, enter payment and shipping details and check out. Here, the recommendation services should be focused on identifying cross-sell/up-sell opportunities to increase the revenue for the site. Some of the services that can be offered are: (1) related product recommendations, i.e., offering recommendations based on the products in the basket or shopping cart; and (2) reminder services, i.e., recommending products the customer buys frequently. The e-commerce site can pop up a screen saying "Don't forget to buy" just before the customer checks out of the site.

By offering such value-added recommendation services at each phase of the online shopping cycle, e-commerce sites can increase the revenue for their site by: (1) inducing their browsers to stay for long and eventually, convert them to buyers; and (2) offer buyers with cross-sell/up-sell opportunities. Current recommender systems use only a few of these services for recommendations. However, a gamut of recommendation services need to be offered to enrich customers’ experiences while browsing/buying and thereby increase the revenue for this site.

**2.5.4 Evaluation of Recommender System**

The evaluation measures for recommender system [30] vary depending upon the objective of the system, i.e., whether the system is a prediction problem or a Top-N recommendation problem. The prediction problem is aimed at identifying how much the user will like a particular item whereas the Top-N recommendation problem identifies a list of products that are of interest to the user. A good recommender system (irrespective of the objective of recommendation) should try to increase the overall quality of recommendations. Here we describe some of the evaluation metrics used in the literature. Figure 2.17 gives the different combinations of recommendations that can be generated in a typical prediction problem. A good prediction system should try to reduce the incidence of false positive and false negative recommendations. The former is the case when products that are disliked by customers are recommended and the latter is the case when products that are liked by customers are not recommended by the system.

There are a variety of evaluation metrics available in the literature, viz. precision, recall, F1 measure, coverage, mean absolute error (MAE), root mean squared error (RMSE), and receiver operating characteristics (ROC). Here we give the definitions of each of these metrics.

|  |  |
| --- | --- |
| **Customer Likes** | **Customer Dislikes** |
| **Recommend** | Found | False Positives |
| **Do not recommend** | False Negatives | Correctly Rejected |

**Figure 2.17:** Possible Recommendations

* Precision measures the degree of accuracy of recommendations produced by the system, i.e., from Figure 2.17.

**Precision = Found/ (Found + False positives)**

* Recall measures the degree of relevant recommendations to the total number of recommendations, i.e.,

**Recall = Found/ (Found + False negatives)**

Alternatively, we can define recall (for Top-N problem) as follows: taking number of hits as the number of items in the test set that were also present in the Top-N recommended items returned for each customer and n as the total number of customers, recall is defined as-

**Recall = Number of hits/n**

* F1 measure is the harmonic mean of precision and recall, i.e.,

**F1 measure = 2 x Precision x Recall/ (Precision + Recall)**

* Coverage measures the ability of recommendation system to produce recommendations for all the items or customers, i.e.,

**Coverage = Number of items (or) customers for which recommendation is**

**Offered / Total number of items (or) customers evaluated**

* MAE is defined as the average of the absolute error; Absolute error is computed as the difference between the rate given by the user and the prediction. That is, given the user ratings, [r1, r2,…,rn], and prediction set, {p1,p2,…pn) where n is the number of items,

**Absolute error, E = {e1, e2,…,en}= {p1-r1, p2-r2,…,pn-rn}**

and,



The prediction problems try to minimize the MAE.

* RMSE is similar to MAE and is biased to provide more weights to larger errors. That is,



* ROC is used in signal detection-systems. It is a plot of the systems sensitivity and (1-specificity), where sensitivity is the probability of a randomly selected good item being recommended by the system and specificity is the probability of a randomly selected item being refused by the system. Two systems are compared by the size of the area under the curve. More the size of area under the curve, the better is the performance.

**2.5.5 Classification of Recommender System**

There are wider varieties of recommender systems [28] available in the literature. The diversity of the systems ranges from the use of the techniques for recommendations (e.g., collaborative filtering, data mining, dependency networks), to the applicability of the system (e.g., retailing, movies, music), the recommendation services offered (e.g., personalized, non-personalized), the type of data used (e.g., implicit, explicit), and so on. Hence, a single classification of the entire gamut of recommender systems is not quite adequate. So, we categorize the recommender systems according to various criteria, as in the following subsections.

1. **Classification Based on the Domain of the Recommender System**

Recommender system based on the domain of the system include systems for different application domains, such as news articles, web pages, movies, music, toys, apparels, retailing, DVDs, and so on. Each of these systems has a specific set of requirements in terms of the data availability, techniques employed, and so on. That is, in a movie recommender system, a customer's ratings of movies can be used for making recommendations. However, in retail situations, the use of customer ratings may not be desirable as there are millions of customers and products. The use of explicit customer ratings is generally sparse and may not be able to produce accurate recommendations. So, a suitable strategy is to use the buying frequency of products or frequency of purchase as an implicit indicator for generating recommendations.

1. **Classification Based on the Type of the Product**

The recommendation systems can be classified depending upon which type of products the recommendations are offered for. Figure 2.18 gives a generic product matrix having value/risk of purchase and volume/frequency of purchase as vertical and horizontal axis, respectively. We have marked four quadrants from 1 to 4 on the matrix. As products in quadrants 1 and 2 have low value/risk of purchase, customers often take minimal amount of effort in making a decision for purchase on those products. Let us refer to such products as low-involvement products (LIPs). Products in quadrants 3 and 4 have high value/risk of purchase & hence the customer takes more involvement in purchasing such products. Let us refer to such products as high-involvement products (HIPs).

In the literature low- and high-involvement products have been defined in different ways. According to some authors; LIPs have four characteristics: (1) a relative lack of active information seeking about brands; (2) little comparison among product attributes; (3) perception of similarity among different brands; and (4) no special preference for a particular brand. Some other authors provide a persuasive model for identifying advertising effectiveness on consumer purchase behavior. They provide do-feel-think model for LIPs and think-feel-do model for HIPs. That is, in a do-feel-think model, a customer first purchases the product on impulse, then feels about the product upon usage, and finally develops an attitude about the product by thinking. Conversely, for an HIP, a customer first thinks and learns about the product, then develops a feeling for the product, and finally makes a decision to purchase. Few other authors distinguish four types of consumer buying behavior based on the degree of buyer involvement and the degree of differences among brands. They define HIPs as expensive, infrequent, and risky. LIPs are defined as most low-cost and frequently purchased products .The LIPs in Quadrant 1 of the matrix are purchased less frequently and have low value. So, the products in this quadrant may not be of much significance, especially for offering recommendations. Similarly, the products in Quadrant 4 have high value/risk of purchase and have high volume/frequency of purchase. In practice, the products that can satisfy both the high-value and high-frequency requirements would be smaller in number and hence may not be of significance to recommender systems.So, recommender systems can be classified broadly on the type of products into LIPs (products in Quadrant-2 of Figure 2.18) and HIPs (products in Quadrant 3).

|  |  |  |  |
| --- | --- | --- | --- |
| **Value/Risk**  **of Purchase** | High | 3 | 4 |
| Low | 1 | 2 |
| Low | High |  |
| **Volume/Frequency of**  **Purchase** | |

**Figure 2.18:** Product Matrix

1. **Classification Based on the Type of Recommendation Service**

Another set of classification of recommender systems can be based on whether recommendations are personalized or not, i.e., personalized recommender systems and non-personalized recommender systems. In the former case, the individual customer preferences are taken into consideration while offering recommendations (e.g., MyCDNow service at Amazon, Match Maker .While in the latter case, average customer preferences or ratings are used for generating recommendations. For generating personalized recommendations, a significant amount of customer data is required for accurate recommendations and hence the amount of personalization that can be done for a new customer is negligible. However, certain amount of personalization can be done even for new customers based on their explicitly stated interests and preferences during the registration process.

1. **Classification Based on Specific Recommendation Services Desired**

We discussed a variety of recommendation services that can be offered specifically in online retailing domain. Each of these recommendation services can be generated by use of different means (i.e., use of techniques, type of data, etc.). So, recommender systems can be classified based on the specific recommendation services desired. That is, systems that offer e-store navigation facilities can be grouped under a class of recommender systems. Similarly, systems that offer wish list service can be grouped under a different class.

1. **Classification Based on the Way Recommendations are Offered**

The recommendations can be offered in an ephemeral or persistent manner. When the recommendations are ephemeral, the users' current preferences alone are taken into consideration while recommending products (e.g., MyCDNow service by [www.amazon.com](http://www.amazon.com)). Such preferences are not stored for future recommendations. However, a persistent recommender system remembers users' past clicks and purchase behaviors and analyzes the same for offering better recommendations (e.g., "Customers who bought" service by www.amazon.com). So recommender systems can be classified on the basis of whether the recommendations offered are ephemeral or persistent

1. **Classification Based on the Degree of User Involvement Needed**

Degree of user involvement is another major criterion for the choice of a recommender system. The system requirements range from no user involvement to complete user involvement (manual and auto classification). Examples of systems that require no user involvement include "Customers who bought" (www.amazon.com) and "Movie Matches" (www.movielens.org). Examples of systems that require more user involvement include "Customer Comments" (www.amazon.com) and "Feedback Profile" (ebay.com).

1. **Classification Based on the Type of Data Available**

Recommender systems can be categorized according to the type of data used for generating recommendations. The type of data can be implicit (navigation and purchase histories), explicit (user ratings of products or items), product information (attributes of products, product taxonomy, and so on), and hybrid (combination of implicit, explicit, and/or product information). The type of data used for generating recommendations is also dependent on the domain of the system, type of application, recommendation service desired, and other criteria.

1. **Classification Based on Techniques Used for Recommendation Generation**

Recommender systems can also be categorized according to the techniques used for generating recommendations. These techniques can be collaborative filtering, data mining, information retrieval, Markov processes, dependency networks, and so on.

The foregoing categorization scheme for the diversity of recommendation system can help: (1) researchers in gaining a better perspective for developing newer systems; and (2) practitioners in choosing the appropriate recommender systems that suit their specific application needs.

* 1. **Collaborative Filtering**

Collaborative filtering [27] (also known as social information filtering) is based on the basic principle of finding a subset of users who have similar tastes and preferences to that of the active user, and offering recommendations based on that subset of users. That is, given an active user ’u’ , compute her ‘n’ similar users [u1, u2, …, un] and predict u’s preference based on the preferences of [u1, u2, …, un].

Collaborative filtering works based on the following assumptions:

* Users with similar interest have common preferences and vice versa
* Sufficiently large number of user preferences is available.

The effectiveness of a collaborative filtering technique is primarily dependent on the above two assumptions. That is, if we have a sufficiently large number of customer preferences and users with similar interest share common preferences, collaborative filtering can accurately predict the, preferences for the users. However, in many commercial applications, getting access to large set of customer preferences is a non-trivial task. So, very often collaborative filtering-based methods suffer from sparsity issues. A detailed description of several drawbacks of collaborative filtering-based methods is given in Section 2.6.2.

Collaborative filtering is essentially done in three stages:

(1) First, select the target user, u, for whom the recommendations are to be generated. Then, compute similarities between the target user and other users in the database. Similarities are computed using similarity metrics such as Pearson correlation coefficient, cosine measure, and so on. A discussion of various similarity metrics available in the literature is made in Section 2.6.1.

(2) Select a set of collaborative users, say [u1, u2, …, uk] who share similar tastes and preferences as those of the target user. The subset of users can be selected choosing (a) the k most similar users, or (b) users with similarity values above a certain predefined threshold.

(3) Offer as recommendation the products liked by collaborative users. A suitable prediction function is used for predicting the likeliness of selected items. The prediction function can be either dependent or independent of the similarity function used. However, the common approach is to use a set of related functions. A simple prediction function used in collaborative filtering is

Where;

is the prediction for the user a on item j ,

ri,j is the rating for user i on item j,

r is the average rating for the user i on all the items,

Sim (a,i) is the similarity value for user a and user i ,

m is the normalizing factor such that the absolute values of the weights sum to unity, k is the number of similar users computed in step (2) above.

An illustration for collaborative filtering process

Let us take an example of a web page recommender system for illustrating the process of collaborative filtering. Table 2.1 gives 10 customers' ratings (on a scale of 1-7) of eight web pages that they have viewed in the past. The pages that are not viewed by them are not rated and hence are left blank in Table 2.1. Now the collaborative filtering process should be able to generate recommendations for web pages that were not viewed in the past by the target customer.

**Step 1:**

In the first step of collaborative filtering process, we have to select a target user for whom web pages are to be recommended. Let us say, we are interested in offering recommendations to the customer C3. A cursory look at Table 2.6 reveals that customer C3 has viewed web pages I1, I3, I4, I5, I6 and I8. So, the recommender system that uses collaborative filtering technique should ideally recommend non-viewed pages, i.e., recommend either the web page I2 or I7. Let us systematically evaluate which of these two web pages would be of interest to the target user.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **I1** | **I2** | **I3** | **I4** | **I5** | **I6** | **I7** | **I8** |
| **C1** | 4 | 2 |  | 3 |  | 6 |  | 2 |
|  |  |  |  |  |  |  |  |  |
| **C2** | 2 |  | 4 | 3 |  | 2 | 1 | 1 |
|  |  |  |  |  |  |  |  |  |
| C3 | 1 |  | 2 | 1 | 2 | 6 |  | 2 |
|  |  |  |  |  |  |  |  |  |
| **C4** | 3 | 5 | 2 |  | 2 |  | 3 | 4 |
|  |  |  |  |  |  |  |  |  |
| **C5** | 1 | 4 |  | 4 | 6 | 5 |  | 2 |
|  |  |  |  |  |  |  |  |  |
| **C6** | 4 | 3 |  | 2 | 4 |  |  | 2 |
|  |  |  |  |  |  |  |  |  |
| **C7** | 5 | 1 | 6 | 2 |  |  | 1 | 1 |
|  |  |  |  |  |  |  |  |  |
| **C8** | 1 | 2 | 7 | 4 |  | 3 |  |  |
|  |  |  |  |  |  |  |  |  |
| **C9** |  | 5 | 1 |  | 3 | 4 | 4 | 2 |
|  |  |  |  |  |  |  |  |  |
| **C10** | 7 |  | 1 | 6 | 1 | 7 | 5 | 4 |
|  |  |  |  |  |  |  |  |  |

**Table 2.6:** Customer Ratings of Web Pages on a Scale of 1-7

**Step2:**

Now, we compute the similarity of customer C3 to every other customer in the database using a similarity metric. For illustration purpose, let us use Pearson correlation coefficient for computing similarity values. The similarity values for Pearson correlation coefficient can range between -1 to +1. A value of -1 indicates that the two users have diametrically opposite preferences whereas a value of +1 indicates that the two users have identical tastes and preferences. For the example considered above, the similarity values are provided in Table 2.7; That is, the table gives the Pearson's correlation coefficient computed between the target user C3 and all other users in the database. Note that Table 2.7 excludes customer C3 as it is the target user.

|  |  |
| --- | --- |
| **Similarity** | **Values** |
| C1 | 0.5958 |
| C2 | 0.2259 |
| C4 | -0.6034 |
| C5 | 0.4042 |
| C6 | -0.3576 |
| C7 | -0.2310 |
| C8 | 0.2193 |
| C9 | 0.1071 |
| C10 | 0.3348 |

**Table 2.7:** Similarity Values Computed for the Target User C3

In the second step of collaborative filtering process, a number of similar (or collaborative) users need to be identified. This can be achieved in two ways:

(1) Select a predetermined number of similar users.

(2) Select set of users with similarity values above a certain threshold.

For example, if we choose option (1) and wish to select two collaborative users, we will select customer C1 (0.5958) and C5 (0.4042), as they have higher positive similarity values (note that we exclude the negative similarity values, as we are interested only in finding users with similar tastes and preferences). On the other hand, if we choose option (2) and set a similarity threshold value as 0.3, we will select the customers C1 (0.5958), C5 (0.4042), and C10 (0.3348) as collaborative users.

**Step 3:**

In the third and final step of collaborative filtering process, we have to offer recommendations to the target user, C3. Initially, we identified from Table 2.6 that customer C3 has not viewed pages I2 and I7. So we will determine the prediction scores for these two web pages and assess which page would be of interest to C3. We use the prediction function outlined earlier in this chapter for computing the prediction scores for items I2 and I7. The weight used in the prediction function is taken as 0.125 (or 1/8). Since there are eight web pages, taking uniform weights for each of these pages will give us individual weights as 0.125 (as cumulative weights should add up to 1). Table 2.8 gives the prediction scores computed for the web pages I2 and I7. Based on the second step of collaborative filtering process described above, the number of similar users selected could be 2 (option 1) or 3 (option 2). From the prediction scores in Table 2.8, we can infer that web page I2 would be of more interest to customer C3 than web page I7. So, web page I2 can be offered as recommendation to the target customer, C3.

|  |  |  |
| --- | --- | --- |
| **Similar Users** | **Web Page** | **Prediction Scores** |
| C1 and C5 | I2 | 1.804 |
|  | I7 | 1.453 |
| C1, C5 and C10 | I2 | 1.642 |
|  | I7 | 1.499 |

**Table 2.8:** Prediction Scores for Web Pages I2 and I7

The foregoing illustration gives a bird's-eye view of how a collaborative filtering-based recommender system generates recommendations. In collaborative filtering process, similar users are identified using similarity metrics. The basic premise is that if the similarity metric has identified people with similar tastes, the chances are high that the target user who belongs to the same group of users will appreciate the recommendations generated.

It can be evident that the choice of similarity metric can influence the final recommendations generated in a collaborative filtering process. So, the similarity metric needs to be judiciously selected. In the following pages, we discuss various similarity metrics proposed in the literature and discuss how they work.

**2.6.1 Similarity Metrics**

There is a variety of similarity metrics [29] available in the literature. Some of the most commonly used measures include Pearson correlation coefficient, Cosine measure, Distance measure(s), Jaccard coefficient, and Tversky coefficient.

**2.6.1.1 Pearson Correlation Coefficient**

Consider two users ‘a’ and ‘b’ who rate a set of ‘m’ items. For example, the ratings of user ‘a’ on item ‘i’ are denoted as ra,i. Pearson correlation between the user ‘a’ and user ‘b’ is defined as the ratio of covariance of user ‘a’ and ‘b’ to that of the product of standard deviation of users ‘a’ and ‘b’. Mathematically,

Where;

ri,j is the user i’s rating for product j , m is the total number of items or products; and

= is the average rating for user x on all the m items.

The correlation coefficient value ranges between -1 and +1. The correlation value of 1 (and -1) is treated as positive (and negative) preferences between the users. A correlation value of 0 means that the users have no common set of preferences. As the correlation coefficients are independent of the mean ratings, users with higher ratings do not affect the overall behavior of the function. The function has been extensively used in collaborative recommender systems for identification of similar users.

An Illustration

Let us say that two customers' preference (indicated on a scale of 1-7) matrices are as follows:

C1 = {1, 3, 5, 5, 7, 3, 2, 0, 4}

C2 = {2, 4, 4, 4, 2, 1, 0, 0, 1}

The Pearson correlation coefficient between these two customers, computed based on the above formula, is 0.5188. If the customer preferences are

C1 = {5, 5, 5, 5, 0, 0, 0, 0}

C2 = {5, S, 5, 5, 0, 0, 0, 0}

C3 = {0, 0, 0, 0, 5, 5, 5, 5}

It can be verified that the correlation coefficient between C1 and C2 is +1 and that between C1 or C2 and C3 is -1. That is, C1 and C2 have identical set of preferences, whereas C3 has a diametrically opposite set of preferences.

**2.6.1.2 Cosine Measure**

Given two vectors and consisting of user ratings, the similarity is defined as the cosine of the angle between the two vectors. That is,

Where "." denotes the vector dot-product operation and the denominator is the square root of the sum of squares of the elements of the vectors (, ).

The cosine similarity between two users is higher if the two users have purchased a larger set of common items. This is especially true for binary-valued data representation, i.e., the vectors are in 0s (non-purchase of an item) and 1s (purchase of an item).

An Illustration

If customer preference vectors are as follows:

C1 = {1, 3, 5, 5, 7, 3, 2, 0, 4}

C2 = {2, 4, 4, 4, 2, 1, 0, 0, 1}

Cosine of the above two customer preferences is computed as-

which is equal to 0.838. If the customer preference matrix is in binary form (i.e., 0 to represent the non-purchase of an item and 1 to represent the purchase of an item),

C1= {1, 1, 1, 1, 0, 0, 0, 1, 1}

C2 = {0, 1, 1, 0, 1, 1, 0, 0, 1}

The cosine similarity metric for the above two customer preferences is 0.5477. It can be easily verified that the cosine similarity value will be higher if the co-purchase of items is larger (i.e., the two customers have purchased a larger set of items in common).

**2.6.1.3 Distance measure(s)**

As per Minkowski distance measure, the distance between two users a and b is defined as

**Distance**

where ra,i is the rating of user a for item i and m is the total number of items. When p = 1 the measure is called Manhattan distance measure, and when p = 2 it is referred as Euclidean distance measure. The similarity between the users is computed using the formula

**Sim (a, b) = 1/[1 + Distance (a, b)] ,Sim (a, b) e** **[0, 1]**

Euclidean similarity (the most commonly used metric) is computed by using the Euclidean distance measure.

An Illustration

If customer preference vectors are as follows:

C1 = {1, 3, 5, 5, 7, 3, 2, 0, 4}

C2 = {2, 4, 4, 4, 2, 1, 0, 0, 1}

then the Euclidean distance computed can be verified as 6.7823. So the Euclidean similarity value is (1/1 + Distance (C1; C2)) = 1/(1 + 6.7823), which turns out to be 0.1285.

|  |  |
| --- | --- |
| **User B** | |
| 0 | 1 |
| **User A** | 1 | a | c |
| 0 | d | b |

a - Count of 1 s for User A alone

b - Count of 1s for User B alone

c - Count of number of common 1s for User A and User B

d - Count of number of common 0s for User A and User B

**Figure 2.19:** Binary Values for Users A and B

From a cursory look at the above illustrations, it can be evident that the same set of customer preferences has given distinctly different similarity values. That is, for the customer preference matrices C1 = {1, 3, 5, 5, 7, 3, 2, 0, 4} and C2 = {2, 4, 4, 4,-2, 1, 0, 0, 1} the similarity values for correlation, cosine, and Euclidean measures were 0.5188, 0.838 and 0.1285, respectively. However, it should be noted that we cannot compare the similarity values generated by different metrics. For interpretation, the similarity values should only be compared with similar values computed using the same metric on other preference matrices.

**2.6.1.4 Jaccard (and Tanimoto) Coefficient**

The Jaccard coefficient [29] measures the ratio of the number of shared attributes. It is commonly used on binary-valued data. It is defined as JacSim(A, B) = c/(a + b + c). When the number of items not purchased by two customers is more (i.e., when d is large), the fractional match coefficient will become very small even though the two customers may have purchased a larger set of common items. So, Jaccard coefficient overcomes this limitation by excluding the value of d in similarity computations.

An Illustration

Case 1:

C1 = {1, 1, 1, 1, 0, 0, 0, 1, 1}

C2 = {0, 1, 1, 0, 1, 1, 0, 0, 1}

Case 2:

C1 = (l, 1, 1, 1,0,0,0, 1, 1, 0}

C2 = {0, 1,1,0,1,1,0,0, 1, 0}

For both the above cases, it can be verified that the Jaccard similarity coefficient is 3/14. So, the Jaccard coefficient, unlike the fractional match coefficient, is not influenced by the number of nonrated items in the preference matrices.

The Jaccard coefficient between two sets A and B is also measured as :

**|AB||AB|**

Tanimoto coefficient is an extension of Jaccard coefficient and is defined as:

**(A.B) / (|A|2 +|B|2-A.B)**

**2.6.1.5 Tversky Coefficient**

Tversky proposed a contrast model in which different weights are given to each of the objects. That is,

**Tversky Sim(A, B) = c / (αa + βb + c)**

where α, β > 0. When α = β = 1, Tversky coefficient is equal to Jaccard coefficient. If we set α= β= 0.5, the coefficient becomes dice index.

When deciding the similarity metric to use, we may have to answer a set of questions: (1) Are we interested in merely assessing the similarity of User A and User B? (2) Are we interested in identifying how similar is User A to User B? The first question is more applicable to clustering and the second one to similar user identification. The α and β values can be set as equal in the case of clustering applications. However, when we are interested in identifying how similar is User A to User B, it is more appropriate to choose a value for α which is greater than that for β. In the extreme case, β can be set to 0. Similarly, how similar is User B to User A can be computed by setting β > α (α = 0 at extreme cases).

An Illustration

Let the customer preference matrices be

C1 = {1, 1, 0, 0, 0, 0, 0, 1, 1} and

C2 = {0, 1, 1, 0, 1, 1, 0, 0, 1}

That is, a = 4, b = 5, c = 1; then dice index is computed as 1/ (0.5 x 4 / 0.5 x 5 + 1) = 0.182.

**2.6.1.6 Fractional Match Coefficient**

Consider n users whose purchases on m items are represented in an n x m matrix. The value for the ith user in the jth column is 1, if user i has purchased item j, otherwise 0. Thus, every user i has a set of binary ratings of his purchases on m items. Now, for any two users A and B, the analysis of their binary ratings can be summarized as shown in Figure 2.19.

Fractional match coefficient between User A and User B is defined as :

**FractSim (A, B) = c/(a + b + c + d)**

That is, it is the ratio of the number of common ratings for users A and B to that of the total count of 1s and 0s for users A and B. So, the similarity between two users would be higher if they rate more items in common. However, this similarity coefficient is strongly influenced by the denominator that includes even the nonrated items (i.e. d).

An illustration

If the preference ratings of two customers are

C1= {1, 1, 1, 1, 0, 0, 0, 1, 1} and

C2 = {0, 1, 1, 0, 1, 1, 0, 0, 1},

Then

a - count of 1s for user C1 = 6;

b = count of 1s for user C2 = 5;

c = count of common 1s for users C1 and C2 = 3; and

d = count of common 0s for users C1 and C2 = 1

So fractional match coefficient can be computed as c/(a + b + c + d), which is 3/15 = 0.2.

If the customer preference ratings were

C1 = {1, 1, 1, 1, 0, 0, 0, 1, 1, 0} and

C2 = {0, 1, 1, 0, 1, 1, 0, 0, 1, 0}

That is the matrix is the same as the earlier one except that it has one additional non-rated item (or item with rating of 0), the fractional match coefficient is 3/16 = 0.1875. As is evident, the similarity values have dropped to 0.1875 from 0.2 even though the number of common items purchased in both the cases is the same. This is due to the additional common non-purchase item in the second case, which has led to a decrease in the similarity values.

**2.6.1.7 Compressed Bit Similarity (cbit)**

A new similarity metric—which can be used for sparse binary data—is given by Srikumar and Bhasker . For using this similarity metric, first the sparse binary user-item matrix is compressed using bitmaps (taking word size, wsz, of the computer as 32 bits, every 32 bits are compressed and stored as an equivalent decimal number). So, if a user U has his item matrix as Ui = {ai1,ai2,…,aim} then compressed form of the item matrix for the same user would be CUi = [bi1, bi2, ..., bisz}, where sz = m/wsz (rounded off to next higher integer). Now, the compressed bit (in short, cbit) similarity between two customers (Ui and Uj) is defined as-

**CbinSim(Ui,Uj)=Σk=1szXk**

Where Xk=0, if (bik=0 or bjk=0) , otherwise Xk=min(bik,bjk)/max(bik,bjk)

So, if a user U1 has purchased all the m items and user U2 has purchased none of the items, their similarity as per the above formula would be 0. Similarly, if U1 and U2 both have purchased all the m items, their similarity would be 1. Different values of similarity between the range of 0 and 1 are obtained for other purchasing patterns between the users.

Unlike other similarity metrics that rely purely on the input data for similarity computations, cbit similarity takes into account the inherent dependencies that may exist between the items. For example, if the item matrices of users A and B are a = {1,0,0,0} and b = {0,0,0,1}, similarity coefficients using current similarity metrics (cosine, Pearson, Jaccard, Tversky) would yield no relationships between User A and User B as they have not purchased any items in common. But, cbit (using wsz = 4) will yield a similarity coefficient of 1/8, i.e., 0.125. This is possible especially for applications like retailing where all the products have hierarchical taxonomy relationships. All the products having the same parent have closer relationship and as we move farther apart in the taxonomy, the relationship diminishes. For instance, items like {Tooth Pastes, Tooth Brushes, Mouth Washes} have some relationship between them (all are dental care products). So, a customer who has purchased only Tooth Pastes and another customer who has purchased only Tooth Brushes do have certain relationships, even though current similarity metrics ignore this. Understanding this relationship can help in identifying suitable products in the product category for recommendations. Though the relationship between the products may not be as simple as they try to capture using compressed bit maps, the metric does help in capturing certain important relationships between hierarchies of products. In addition, the metric is highly scalable compared to other similarity metrics.

An Illustration

Let us take the following products: {Tooth Pastes, Tooth Brushes, Mouth Washes, Low fat milk, Medium fat milk, and Heavy fat milk}

Let us assume that the customer purchase matrices (on the above products) are as follows:

C1 = {1, 0, 0, 0, 1, 0} and

C2 = {0, 0, 1, 1, 0, 1}

A cursory look at the above preference matrices would reveal that similarity metrics such as cosine and Jaccard would give a similarity value of 0, indicating that there is no relationship between the purchase matrices of customers C1 and C2. On the other hand, if we compute the similarity at a higher level of the taxonomy, then the customer preference matrix can be transformed as-

C1 = {1, 1} and C2 = {1,1}

Now, there are only two categories, namely dental care & milk products. As each of the customers has purchased at least one product in each of the category, the transformed matrix will have an entry of one against each product category. Using cosine metric on this transformed matrix would give a similarity value of 1, indicating that both the customers have identical tastes & preferences. But, we know the implicit relationships that exist between the products (shown as taxonomy in Figure 2.20). Cbit essentially tries to capture this implicit relationship using a metric.

**Figure 2.20:** Sample Product Taxonomy for Consumer Products

As per the cbit metric, taking the word size as 3 since the first two and last three products belong to a particular category (note that the word size can be variable based on the product groupings derived), the transformed value of the preference matrix is

C1 = {100, 010} and

C2 = {001, 101} which is represented in compressed bit form as

C1 = {4, 2}

C2 = {1, 5}

i.e., 001 is converted to its equivalent decimal form and represented as 1. Similarly, 101 is converted to 5, which is its decimal form. Now, cbit(C1, C2) = min(l, 4)/max(1, 4) + min(2, 5)/max(2, 5), which is 1/ 4+ 2/5 = 0.25+ 0.4 = 0.65.

**2.6.2 Drawbacks of Collaborative Filtering**

Collaborative filtering suffers from the following drawbacks, namely:

* First rater problem
* Sparsity
* Scalability

We will discuss each of these issues at greater length in the following subsections.

* **First Rater Problem**

Collaborative filtering-based systems require a large set of customer preferences for predicting the new preferences accurately. That is, for a reliable collaborative filtering-based prediction system, we need to have a large set of customer ratings on a variety of items available on the database. As the system can predict accurately only after it has gathered a large set of opinions, customers will not be willing to express their detailed preferences in the initial stages (if it is not going to help them in any way, despite offering preference ratings). This first rater or cold start problem can be overcome by gathering implicit preferences. The implicit preferences could be customer's past purchases (indicating the customer's preference over the products or items that he has not purchased), web sites consulted by the user, or click histories captured during customer's navigation in the web site.

* **Sparsity in Collaborative Filtering**

Collaborative filtering-based methods require customers' subjective ratings of products or items. However, the numbers of ratings obtained are usually much smaller compared to the number of items, leading to the problem of sparsity. Even in retail applications, where the buying frequency of products or the number of times a product is purchased has been used as an implicit rating ,the problem of sparsity is inevitable. This is due to the fact that customers buy far fewer products compared to thousands of products available in the retail site. Some of the solutions offered in the literature to address these issues are: (1) the use of singular value decomposition (SVD) techniques to extract the reduced set of components that minimizes the total variance; and (2) use of product categories rather than brand names or stock-keeping units (SKUs) of products.

* **Scalability Issues**

Collaborative filtering-based systems require computation of similarity among customers. That is, if there are 100 customers in a database, then the similarity has to be computed for 99C2 cases, which is of the order of n2. So, if the number of customers in database is large, the computation time becomes significantly greater, leading to the problem of scalability. In real-life environments, where there are large numbers of customers and products, providing real-time performance becomes critical.

* 1. **Chapter Summary**

This chapter provided the prominent & relevant research that has been undertaken related to the proposed approach. The chapter discussed the evolution of web and various services and applications being offered by Web 2.0 providing an overview of virtual community in a Web 2.0 world and whether the recent social networking sites follow a truly collaborative approach or a simply participatory one. It also provided a detail study of recommender systems followed by the study about the collaborative filtering technique. The next chapter illustrates the novel techniques that constitute the proposed approach to address the issues presented in this chapter.