InnoJam

A Web 2.0 discussion platform featuring a recommender system

Miquel Sonsona Villalobos

Knowledge Engineering and Machine Learning Group

Artificial Intelligence Master Thesis

Departament de Llenguatges i Sistemes Informàtics

Universitat Politècnica de Catalunya

Advisor:

Ulises Cortés

Barcelona, January 2010

Abstract

In this Master Thesis we have designed, implemented and evaluated a Web 2.0 platform for massive online-discussion, inspired by Innovation Jams.

Innovation Jams, the original initiative from IBM, has proven to be successful at bringing together vast amounts of people, capturing their untapped knowledge and, while the participants are discussing, gather useful insights for a company's innovation strategy [Spangler *et al.* 2006, Bjelland and Chapman Wood 2008].

Our approach, based in an open-source forum system, features visualization techniques and a recommender system in order to provide the participants in the Jam with useful insights and interesting discussion recommendations for an improved participation.

A theoretical introduction and a state-of-the-art survey in recommender systems has been gathered in order to frame and support the design of the hybrid recommender system [Burke 2002], composed by a content-based and a collaborative filtering recommenders, developed for InnoJam.

Acknowledgments

First of all, I would like to thank my future wife, Cristina, and my parents and sister, my family, for always being there to support me. With them, grandparents, aunts, uncles and cousins have help me grow into who I am; those who are, and also those who were, I'll always bring you with me.

Secondly, I would like to thank my fiancée's family, for teaching me to learn to live against difficulties. Life's not always easy, and it's hard to cope with that. We all should value things in their fair measure.

Thirdly, I would like to thank all my friends for having such great times together. The list is long and most of them will not read this.

Moreover, I would like to thank all the members of KEMLg and specially the ones who lived in the old Ω-S206 and the ones who live in the new K2M-201, for always serving as inspiration and making our work environment a warm and lovely place.

More in particular, I would like to thank specially Arturo, Luis and Agustí, for sharing countless hours of work and fun, as well as coffee time and everything that entails.

My colleagues in the projects I have collaborated with also deserve my gratitude for working hard and having fun in the meantime. A special thought goes to Fran and his coffee machine in Castelldefels. Here again, coffee is the excuse.

A special and warm "*thank you*" to the people in the department's administration, who always have been nice and helpful when any need urged.

Another special and warm "*thank you*" to Keith, for his help making me resemble a person who writes and speaks English well.

Last, but not least, many thanks to Ulises and Esteve for giving me the opportunity to work with them and for guiding and advising me during the development of this thesis.

Thanks all.

Table of Contents

1. Introduction

Since ancient civilizations like the Greeks, people's exchange and development of ideas and opinions have been made by means of discussing with other individuals or groups of people about several topics.

In the Internet era, this information exchange has evolved to the next level. It can be done either synchronous or asynchronous, either anonymous, private or public, one-to-one or many-to-many. Via email, chat, videoconference, voice... the possibilities are almost unlimited.

Given that fact, large transnational companies started to think on *how* to use IT in order to put their entire workforce into solving common widespread problems. They realized that a platform where thousands, hundreds of thousands of employees could share and exchange knowledge was needed. This is *how* innovation jams emerged.

Innovation Jams are events that draw together vast groups of users, usually globally widespread, to come up with new ideas and decisions about a collection of predefined topics. The first organization to implement such a crowd-sourcing solution was IBM with background research in the field of online conversations during the late 1990s and putting it into practice since the beginning of 2000s.

> *“The largest online brainstorming session ever.”*
>
> *“Every idea counts.”*
>
> *IBM press release*

Since 2001, IBM has used jams to involve its more than 300,000 employees around the world in far-reaching exploration and problem-solving. ValuesJam in 2003 gave IBM's workforce the opportunity to redefine the core IBM values for the first time in nearly 100 years. During IBM's 2006 Innovation Jam - the largest IBM online brainstorming session ever held - IBM brought together more than 150,000 people from 104 countries and 67 companies. As a result, 10 new IBM businesses were launched with total seed investment of $100 million.

Although these first experiences were related to IBM, jams are not only restricted to business. Their methods, tools and technology can also be applied to social issues. In fact, in 2005, over three days, the Government of Canada, UN-HABITAT and IBM hosted Habitat Jam. Tens of thousands of participants - from urban specialists, to government leaders, to residents from cities around the world - discussed issues of urban sustainability. Their ideas shaped the agenda for the UN World Urban Forum, held in June 2006. People from 158 countries registered for the jam and shared their ideas for action to improve the environment, health, safety and quality of life in the world's burgeoning cities.

In the most recent experience, the 2008's InnovationJam tapped employees from more than 1,000 companies which produced 32,000 posts during a 90 hours period, focusing on the 'Enterprise of the future'. Having such a global situation with the financial crisis, the event was a good opportunity to use the system when actual applicable results were really interesting to achieve.

According to [Bjelland and Chapman Wood 2008] the main lessons learned from these first pilots were :

- Many people throughout an organization may have important strategic ideas.
- Online conversations and sophisticated technology can bring those ideas to bear on important societal problems and make them worth millions to a company.
- Limitations on how most people recognize and build on others' ideas online
- Jams can significantly speed the path to decision and action.

The last event to be held in collaboration with the Brussels-based think tank Security & Defence Agenda[1] will build learning and consensus on issues of security and defense. SecurityJam[2] will gather thousands of subject-matter experts and other thought leaders from business, government and nongovernmental organizations to analyze and clarify new threats to international peace and security. The result will be a set of recommendations which will go to the leadership of the E.U. and NATO in April 2010.

[1] Website at *http://www.securitydefenceagenda.org/* (last visit on January 4, 2010)

[2] Website at *http://www.securitydefenceagenda.org/SecurityJamSession/tabid/967/Default.aspx* (last visit on January 4, 2010)

1.1. Motivation

Aiming at bringing together a *massive* number of people discussing leads to several implications on what has to be considered and addressed in order to achieve some success within this endeavor.

When trying to extract some actionable knowledge from a massive online discussion, some sort of technique needs to be used in order to funnel all knowledge provided by the participants into ideas, proposals, actions, etcetera, which can be directly applied or taken into account by decision-makers.

In the IBM's approach, the techniques used were text mining and text summarization complemented with a dedicated team of moderators, who supervised the classification of discussions into categories [Spangler, *et al.* 2006].

In our approach, the techniques we provide are useful visual and textual information about the event and a recommendation engine to tap the overload of contributions generated by the community of participants [Sonsona and Almirall 2009].

Recommender systems have been an important line of research, in Artificial Intelligence, since the 1990s [Goldberg *et al.* 1992, Resnick *et al.* 1994, Shardanand and Maes 1995], and have constituted a solution for users when dealing with vast amounts of information. It is also a common technique in eCommerce companies, like Amazon [Linden *et al.* 2003, Leino and Räihä 2007] or Netflix, which has become very popular since the launching of its $1M contest[3] to improve its movie recommender system [Bennett and Lanning 2007]. Furthermore, the European Commission is funding projects that use and research recommender systems, like MyMedia[4], in collaboration with Microsoft.

Some of the use cases where a recommender system is considered to be useful are [Herlocker *et al.* 2004]:

- Filtering information to provide only worth-consuming information, predicting and distinguishing between desired and undesired content.

---

[3] Website at *http://www.netflixprize.com/* (last visit on January 4, 2010)

[4] Website at *http://www.mymediaproject.org/* (last visit on January 4, 2010)

- Finding good items among all those present in the system. There exist special cases where users want or need to reduce the amount of not-good-enough items.
- Experimenting with the system, just for pleasure of doing so, or to test whether the recommender system is able to capture user preferences correctly or not.
- Social uses. Users use the recommender system to contribute to its improvement or to benefit from the community, even influencing and introducing bias into the system.

In the recent IBM Global CIO Study [IBM 2009], business intelligence and analytics, understood as the ability to see patterns in vast amounts of data and extract actionable insights, were identified as the most promising way to enhance organizations' competitiveness and ability to meet client needs. In line with this, visualization is the forefront aspect of these much valued techniques. The European Commission is also funding several projects in visualization research like QVIZ[5], VIDI[6] or WAVE[7].

Moreover, the use of innovation jams and other Web 2.0 and Web 3.0 tools, like social networks, mass-collaboration platforms, prediction markets, multi-lingual and semantic interoperability, visualization techniques, etcetera, has been recommended by the European Commission in the eParticipation field [Millard 2009].

1.2. Objectives

Based on all these ideas, the main objectives we want to achieve with this thesis are:

- The design and development of an online discussion platform.
- Use and develop visualization techniques in order to provide a better understanding of the event to the participants.
- Use and develop a recommender system able to provide users with recommended discussions.

1.3. Related work

Old days in the Internet where companies published content and users were just mere information consumers are over. With the advent of technologies like blogs,

---

[5] Website at *http://qviz.eu/* (last visit on January 4, 2010)

[6] Website at *http://www.vidi-project.eu/* (last visit on January 4, 2010)

[7] Website at *http://www.wave-project.eu/* (last visit on January 4, 2010)

microblogs and wikis, the decrease of technical requirements both for the users and for their machine for publishing and more recently the burgeoning of social networks, the game has changed, it's time for *prosumers* [Tappscott and Williams 2006].

> *"Web 2.0 is the design of systems that harness network effects to get better the more people use them."*
>
> *Tim O'Reilly*

In the Web 2.0 [O'Reilly 2005] era, where large groups of people are supposedly smarter than even an elite few, no matter how brilliant—better at solving problems, fostering innovation, coming to wise decisions, even predicting the future [Surowiecki 2004], crowdsourcing and open collaboration projects flourish on the net in a myriad of flavors: music and videos, bookmarks, encyclopedias, product reviews... there is even a wiki playbook about this wiki movement[8], and a website[9] to crowdsource any task one could imagine [Hoffmann 2009].

> *"A powerful global conversation has begun. Through the Internet, people are discovering and inventing new ways to share relevant knowledge with blinding speed. As a direct result, markets are getting smarter—and getting smarter faster than most companies."*
>
> *The Cluetrain Manifesto*

Some of them are more focused on corporative environments, and companies have applied Innovation Jams in their R&D activity or have developed their own solutions for tapping innovation and creative knowledge from within their own employees like Nokia's

[8] Wikinomics playbook is a book written by the community. It can be found online at *http://www.socialtext.net/wikinomics/index.cgi* (last visit on January 4, 2010)

[9] The Mechanical Turk is a community of crowdsourcers owned by Amazon, where participants get paid for doing commanded "*Human Intelligence Tasks*". Website at *https://www.mturk.com/mturk/welcome* (last visit on January 4, 2010)

BetaLabs[10], Intel's IT Galaxy UK[11] and Procter&Gamble's Connect+Develop[12] [Howe 2006] or Starbucks idea pool[13] open to customers to have their say.

With these tools, companies have advanced their market positions, improving and expanding their product development and being drawn to a more open model of management, thanks to open collaboration [Gabor 2009]. The case of Procter & Gamble has reported amongst other results, after six years of application, a significant increase in the external origin of products and a more successful, and inexpensive, R&D and innovation initiatives [Procter & Gamble 2006].

But there are also cases some which are being developed in the political and social arena. Websites like essembly[14], hotsoup or MoveOn[15] are promoting several ways of virtual interaction between citizens or between citizens and decision-makers, in order to promote political engagement and allow citizens to find their political voice in a system dominated by big money and big media. There are initiatives of collaborative government in the United States[16] and in New Zealand[17]. Dijksman.com[18] is a company which builds collaborative solutions for municipalities. A very successful application of the open innovation principles is the city of Manor, Texas, with their initiative ManorLabs[19]. Manor Labs is an open innovation platform designed to allow their citizens to help the municipality generate new ideas and solve problems for the local government agency. For more information on *how* this innovative case is implemented, it is worth it to read a recent article on govfresh.com[20]. Another interesting case is presented by The United Nations University, which runs OurWorld2.0[21], a platform for collaborating for a better world.

---

[10] Website at *http://betalabs.nokia.com/* (last visit on January 4, 2010)

[11] Website at *http://itcommunity.intel.co.uk/community/uk* (last visit on January 4, 2010)

[12] Website at *http://www.pgconnectdevelop.com/* (last visit on January 4, 2010)

[13] Website at *http://mystarbucksidea.com* (last visit on January 4, 2010)

[14] Website at *http://www.essembly.com/* (last visit on January 4, 2010)

[15] Website at *http://www.moveon.org/* (last visit on January 4, 2010)

[16] Website by the Obama administration at *http://opengov.ideascale.com/* (last visit on January 4, 2010)

[17] eParticipation New Zealand Government wiki at *http://wiki.participation.e.govt.nz/wiki/Main_Page* (last visit on January 4, 2010)

[18] Website at *http://www.dijksman.com/* (last visit on January 4, 2010)

[19] Website at *http://www.manorlabs.org/* (last visit on January 4, 2010)

[20] Article at *http://govfresh.com/2010/01/whiteboard-innovation-how-manor-ideas-become-solutions/* (last visit on January 4, 2010)

[21] Website at *http://ourworld.unu.edu/en/* (last visit on January 4, 2010)

Summing up, there are lots of initiatives which try to benefit from the wisdom of crowds. However, not many of them are applying any sort of Artificial Intelligence, Data Mining, Business Intelligence, etcetera. techniques to leverage the knowledge they are collecting from their users.

1.4. Organization of this thesis

The rest of this thesis is organized as follows:

- **Chapter 2 - State of the Art.** This chapter collects background knowledge and reviews the state of the art in recommender systems.
- **Chapter 3 - Recommender System Development.** This chapter presents the design and development of the recommender system proposed by this thesis work.
- **Chapter 4 - Platform Development.** This chapter stands for the prototype documentation of InnoJam, the system that represents the practical outcome of this thesis.
- **Chapter 5 - Practical Applications.** This chapter briefly describes how InnoJam has been applied in live cases and in which other scenarios it could be useful.
- **Chapter 6 - Evaluation.** This chapter describes the results of the experiences in the previous chapter.
- **Chapter 7 - Conclusions & Future work.** This chapter provides a recapitulation of the outcomes of this thesis and what are the main conclusions we have obtained from this work. It also outlines research lines and paths for the continuation of the work in InnoJam, or, put it in another way, what could make InnoJam better.
- **References.** The references used during the research and the development of this thesis can be found in this section.

2. State of the Art

As seen in chapter 1, a need arose for Information Systems to be able to offer users with valuable pieces of information in a scenario of information overload. Recommender systems address this issue basically by filtering and predicting the likeliness of this pieces of information with regard to user preferences, which have been expressed either explicitly by the user or implicitly during the use of the system or by other means.

Another practical use for recommender systems is to recommend users to other users. This functionality has mainly two applications:

- team assembly. In business environments, human resources departments may build groups of people in order to make them work together based on their competences [Tejeda *et al.* 2009, McDonald 2003].
- social discovery. In social networks or dating sites, recommenders may be used to allow users to meet new people based on their interests or other criteria, like demographic information [Terveen and McDonald 2005].

In this chapter, the recommendation problem is formally defined and the different approaches and techniques to solve it are described, presenting some factors that affect them and some solutions to these drawbacks.

Throughout the chapter, some considerations and side notes are drawn in relation to the design of the recommender system developed for this thesis work, which is widely described in chapter 3.

2.1. The recommendation problem

The recommendation problem is briefly described as estimating ratings for the items which have not been observed by a user [Adomavicius and Tuzhilin 2005]. However, two subproblems can be distinguished [Sarwar *et al.* 2001, Celma 2008]: first, a prediction problem, and second, a recommendation problem.

2.1.1. The prediction problem

The prediction problem deals with estimating items' likeliness for a given user. This problem is the most relevant in the research on the recommender systems field, because it represents the degree of interest that the user could potentially have for the items to be recommended.

2.1.2. The recommendation problem

The recommendation problem deals with creating the list of recommended items, given that each items has a previously computed likeliness measure. This problem, although being less relevant than the prediction problem, constitutes in fact the critical problem for the user, as it will directly affect how she will perceive the recommendations the system provides. Most of the research in the interaction between the user and the recommender system is conducted to deal with this problem: user interfaces in recommender systems, recommendation explanation, item diversification, user-centric evaluation, user satisfaction, recommendation, effectiveness, etcetera [McNee *et al.* 2006, Tintarev and Masthoff 2007, Leino and Räihä 2007, Rubens and Sugiyama 2007].

2.1.3. Formalization

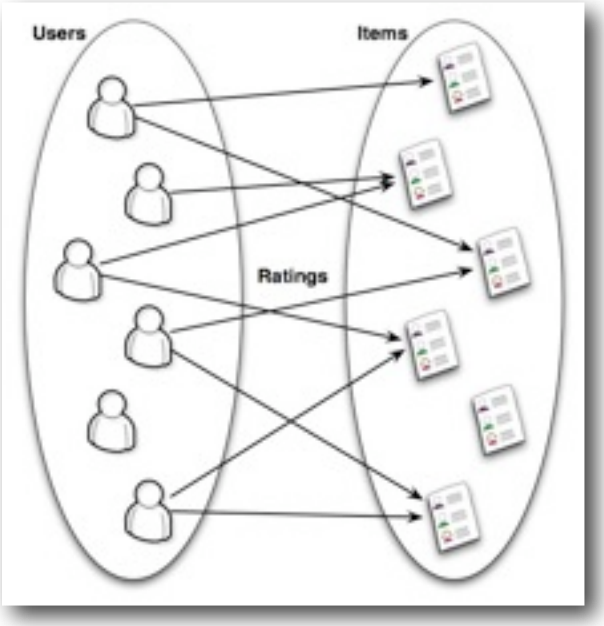
In a typical recommendation scenario, let $U = \{u_1, u_2, .. u_m\}$ the set of users in the system and $I = \{i_1, i_2, ... i_n\}$ the set of items to be recommended.

Each user u_i has a list I_{u_i} of items which the user has expressed her interests in. This set of items is usually a subset of the whole set of available items $I_{u_i} \subseteq I$. It can even be an empty set $I_{u_i} = \emptyset$, especially when the user is new in the system.

Given an active user $u_a \in U$, $P_{a,j}$ expresses the predicted likeliness of item $i_j \notin I_{u_a}$.

After the prediction, a list of N items $I_r \subset I$ is built, with the N items the user will like the most, *i.e.* the N items with the highest likeliness prediction $P_{a,j}$. This list of recommended items should not contain items already observed by the user $I_r \cap I_{u_a} = \emptyset$.

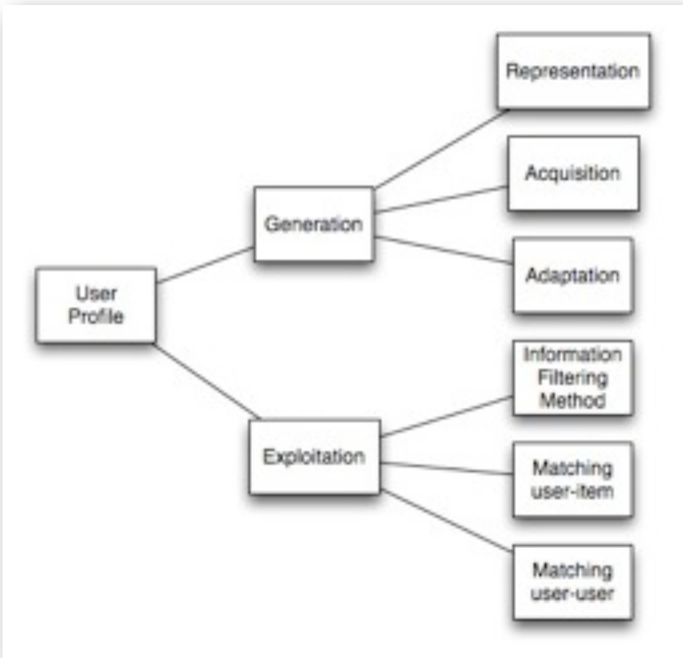
The prediction function is usually represented by ratings, which can be defined as triples $\langle u,i,r\rangle$ being r the value that either implicit or explicitly the user u assigns to item i .



Conceptual scheme of a typical recommendation scenario

2.2. User profile

The general recommendation problem can be described from the user point of view in terms of profile generation and profile exploitation.



The recommendation problem conceptually depicted with respect to the user profile

2.2.1. Generation

The generation of a user profile is subdivided into three main aspects: the representation, the acquisition and the adaptation.

Profile representation

The representation of a user profile contains the information the recommender system requires to model user preferences. These preferences are directly related to what is being recommended, therefore, the information contained in the user profile strongly depends on the domain and the items being recommended.

Although there are several features which can be considered when designing the user model, like the user's knowledge, the user's interests, the user's goal, the user's background or the user's personality; the most used are, by far, the user's interests.

The traditional representation of the user interests is done using weighted keyword vectors. These vectors can be automatically extracted by the system from the items the user is rating.

Profile acquisition

The acquisition of the initial profile of the user can be obtained by several means, which differ in the level of complexity for the system and in the level of effort required by the user:

- **Empty.** This is the simplest approach to the generation of a user profile. The system just creates an empty profile. However the system will not be able to recommend any item to the user as it does not have any information about the user preferences or interests.
- **Manual.** This approach requires the user to input some information. The information the user is required to enter may be that related to the interests, to personal information or the system may obtain user preferences for a predefined set of training examples.
- **Import.** This approach prompts the user to provide an existent profile on other application or external repository. A relevant possibility is to use semantic web representations like FOAF[22] [Codina 2009, Celma 2008].

[22] Website at *http://www.foaf-project.org/*

- **Stereotyping.** This approach classifies the user into one of the predefined stereotypes based on some personal information.

Profile maintenance

For maintaining the user profile, the information collected may be explicitly provided by the user or the system itself implicitly extracts it from the user interaction with the system and user behavior patterns.

Each of these approaches have some drawbacks. If the user has to explicitly provide information it becomes a costly effort and if the user does not provide the information, then the profile becomes inaccurate with respect to the user interests and preferences. However, if the system implicitly captures this information, the system is only able to capture positive feedback, as there is no reliable way to capturing negative feedback without explicitly asking the user to provide it.

Profile adaptation

The adaptation of the user profile is a key concept when intending to model the user interests or preferences even when they evolve or change during time. There are three main approaches to this task:

- **Manual.** This approach requires the users to update their own profile
- **Additive.** This approach automatically adds new information to the profile. It is the most common approach.
- **Forgetting mechanisms.** This approach assumes an evolving nature in user interests and preferences thus applying a gradual balance in the relevance of old information versus newer.

2.2.2. Exploitation

The exploitation of a user profile is also subdivided into three aspects: the information filtering method, *i.e.* the recommendation algorithm, and both the matching mechanism between users and items and between users. Recommendation algorithms and matching mechanisms are described in the next section.

2.3. Recommendation methods

This section describes the most common methods used in recommender systems for filtering the whole space of items susceptible to be recommended [Ramezani et al. 2008].

2.3.1. Demographic filtering

Demographic filtering relies on users' descriptions to classify profiles in clusters, and then learn the relationship between these clusters and items. The filtering of information is reduced to the users of the group which the user belongs to [Krulwich 1997].

Limitations

In this approach, recommendations are usually too general, as they are meant to fit to a group of users or stereotype.

2.3.2. Knowledge-based filtering

Knowledge-based recommenders use domain knowledge to generate recommendations [Tejeda 2006]. Three subtypes can be distinguished depending on which knowledge-based technique do they use for recommending: case-based reasoning, constraint-based reasoning and rule-based systems.

The main benefit for this method is the avoidance of the cold-start problem and the main limitation is the costly process for acquiring and maintaining knowledge required.

2.3.3. Collaborative filtering

Collaborative filtering recommends items based on the preferences of users with similar tastes. The name of this technique comes from the Tapestry project, which firstly introduced it [Goldberg et al. 1992].

Two different approaches can be distinguished based on the element which the similarity is computed by: **user-based** when the recommended items are those preferred by similar users; **item-based** when the recommended items are similar to those the user has preferred in the past, specially in terms of co-rating with other users.

Usually, specially for collaborative filtering recommender systems, a matrix where users and items are respectively represented by rows and columns is used for storing users' preferences. $R_{u,i}$ is the rating of user u for item j .

User-based collaborative filtering

User-based collaborative filtering considers the preferences of similar users when computing the predicted rating of user u for item i , $P_{u,i}$. It can be computed as the mean of ratings for item i of users similar to u :

$$P_{u,i} = \bar{R}_u + \frac{\sum_{v \in Neighbors(u)}^{k} sim(u,v)(R_{v,i} - \bar{R}_v)}{\sum_{v \in Neighbors(u)}^{k} sim(u,v)}$$

Predicted rating equation

where $\bar{R}_u$ is the average rating value of the user u , $Neighbors(u)$ is the set of similar users to u , computed by the similarity measure $sim(u,v)$ and of size k .

The most common approaches for finding a given user neighbors are Pearson correlation and cosine similarity. This similarity measure is essentially a distance measure between users, and represents the weight similar users will have in the predicted rating.

$$sim(i,j) = \frac{Cov(i,j)}{\sigma_i \sigma_j} = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_i)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_j)}^2}$$

Pearson correlation equation

Item-based collaborative filtering

Item-based collaborative filtering considers similar items to the target item. In this case, the similarity is understood in terms of users having rated the same items.

The most common similarity measures are cosine similarity, adjusted cosine similarity and Pearson correlation.

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|i\| * \|j\|} = \frac{\sum_{u \in U} R_{u,i} R_{u,j}}{\sqrt{\sum_{u \in U} R_{u,i}^2}\sqrt{\sum_{u \in U} R_{u,j}^2}}$$

Cosine similarity equation

The adjusted cosine measure, however, is more robust as it takes into account the average of the user's ratings when considering the co-rating of items:

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_u)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_u)^2}}$$

Adjusted cosine similarity equation

After the similarity between items has been computed, the predicted value for the target item is obtained from the values the user has given to its similar items:

$$P_{u,i} = \frac{\sum_{j \in NeighborsRated(i,u)} sim(i,j) R_{u,j}}{\sum_{j \in NeighborsRated(i,u)} sim(i,j)}$$

Predicted value for item i to user u

where $NeighborsRated(i,u)$ is the set of items rated by the user u which are similar to item i.

Limitations

The collaborative-filtering approach suffers when users' ratings are very sparse, specially when performing user-based similarity, because it can be difficult to find reliable neighbors. This problem also affects users with interests different from the average, which will not have many neighbors.

Collaborative filtering is very prone to the cold-start problem both for new users and for new items. A low number of ratings implies that finding similar elements is less reliable.

As being solely based on users' ratings, it captures well the social perceptions of the users with regard to items, but this fact also produces a bias towards the most popular items, which as a result of the similarity computation, will be recommended more usually than less popular items, without considering the fact that this less popular items could be more relevant for the user [Celma 2008].

2.3.4. Content-based filtering

Content-based filtering recommends items based on the results of matching the user interests against features of items similar to those the user liked in the past. Some key points in this approach are the items' descriptions and the similarity measure between items.

Some common distance measures for two features vectors x and y , are the Euclidean distance, the cosine similarity, the adjusted cosine similarity or the Mahalanobis distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Euclidean distance equation

$$d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$$

Mahalanobis distance equation

However, when features are not numeric a delta function can be used for defining a distance. For example, given a delta function like: $\delta(a,b) = 0 \Leftrightarrow a = b$ and $\delta(a,b) = 1$ otherwise, the distance metric would be:

$$d(x,y) = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i, y_i)$$

Content-based filtering solves some of the problems of collaborative filtering methods as the early-rater problem, because and the popularity bias, as there is no human intervention in the process.

Limitations

The main limitation of content-based filtering comes from the fact that it only takes into account the features that describe the items. When two items are described by the same set of features and their values are the same, their differences become imperceptible, and hence the items resemble the same. Moreover, given that, the system is not able to take into account the real interests of the user when recommending.

This approach also suffers the new user problem. As no items have been rated by the user when entering the system, the recommender system does not have information on which items to recommend.

A mechanism for augmenting diversification in the item recommendation should be introduced to the system to overcome the fact that too similar items to those the user liked would always be recommended, thus suffering from overspecialization.

2.3.5. Context-based filtering

Context-based methods use contextual information to describe items, *e.g.* time or place. Usually the contextual information is used to reduce the ratings space to those pertaining to the current context [Adomavicius et al. 2005].

The first attempt at incorporating contextual information into a recommender system [Herlocker and Konstan 2001] showed that certain applications may benefit from the inclusion of knowledge about user's task for obtaining better recommendations.

One of the most interesting techniques is social tagging, which extends the ratings matrix for users and items with a third dimension for tags. This could be explored as future work, see section 7.2.2.

2.3.6. Hybrid approach

Hybrid methods for recommendation combine two or more of the previously described techniques in order to add their strengths and overcome their respective limitations. Usually, collaborative filtering is one of the combined techniques [Basilico and Hofmann 2004].

Two or more approaches have to be combined in order to retain the benefits and discard the drawbacks. The most common combination of methods are [Burke 2002]:

Hybrid method	Description
Weighted	The scores of several recommendation techniques are combined together to produce a single recommendation.
Switching	The system switches between the different combined recommendation techniques depending on the current situation.
Mixed	Recommendations from several different recommenders are presented at the same time.

Hybrid method	Description
Feature combination	Features from different recommendation data sources are thrown together into a single recommendation algorithm.
Cascade	One recommender refines the recommendations given by another.
Feature augmentation	Output from one technique is used as an input feature to another.
Meta-level	The model lear ned by one recommender is used as input to another.

2.4. Factors and problems

Summarizing, the following are the main problems that recommender can face and some of the solutions proposed in the literature.

2.4.1. The cold start

When relying on a collaborative-filtering approach a problem arise whenever a new user or a new item enter the system. In fact, the problem arise when the new element is the one which is used to compute the similarity.

As a new user has not rated any item yet, it is not possible to find similar users based on ratings that does not exist, so no new items can be recommended based on what similar users have rated. The same happens when a new item is introduced into the system, as it has not been rated by any user, it will not be recommended as it will not be similar, in terms of ratings, to any other item rated by the user.

One of the simplest solutions for overcoming the cold start problem when a new user starts using the system is to provide a "*less personalized*" list of recommended items. This list could contain either **a subset of random items or the most popular items, or a combination of both** [Celma 2008]. Research in this area includes smarter selection of items which should be rated by the new user [Kohrs and Merialdo 2001].

2.4.2. Sparsity

A recommender system usually works with large numbers of users and items, but the fact that number of ratings is, in comparison, very low, is a common problem. There could be the case where some items were seldom rated, these items would also be rarely recommended in a collaborative filtering approach.

This problem can be approached by **expanding the information available about users** in their profile to compute the similarity (and thus moving into a hybridized solution closer to demographic filtering) or by using a method for **reducing the dimensionality** of the rating matrix, *e.g.* Singular Value Decomposition [Sarwar *et al.* 2000].

2.4.3. Scalability

Traditionally, Collaborative Filtering recommender systems find their bottleneck in terms of performance in the user similarity measure computation [Sarwar *et al.* 2001]. When the user base is relatively big, this bottleneck privates the system to provide real-time recommendations. One solution to overcome this issue is to **rely in a model-based approach**, isolating the neighborhood generation and the prediction generation steps.

Another solution would go for **reducing the dimensionality**, which would produce a less complex computation in terms of both time and space, thus speeding up the process. There are several ways for reducing the dimensionality, for example using matrix factorization techniques [Sarwar *et al.* 2000] or using existing knowledge about the users like the friends or contacts of a user in a social network context.

3. Recommender System Development

As seen on chapter 2, the task of developing a recommender system involves taking several crucial decisions. Amongst others, the most important are:

- the way in which users profiles and items descriptions are modeled, *i.e.* which features do represent each of the elements in the system.
- how these features are initialized and updated.
- how the similarity measures between users and between items are computed, based on their description.
- which filtering method is used for selecting the items to be recommended.

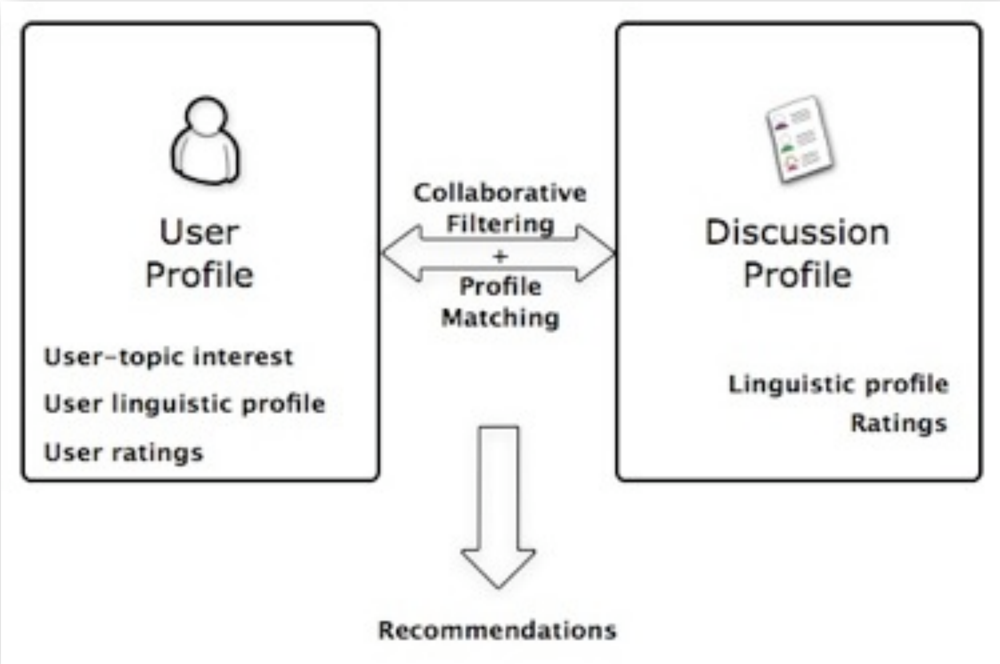
Apart of these questions, which will be described later in this chapter, there are some general factors which affect the design of the recommender system. These factors are inherent in the kind of events (the jams) which we want to apply our system to, and we take it as assumptions for the design of the recommender system. Afterwards a series of other less specific factors affecting the design will be described.

- Jams have a short duration, therefore the use of the system is relatively short and concentrated in this period.
- Jams are structured around different topics.
- Jams are text-based discussions, hence the items to recommend are discussions.
- There is not any knowledge about the users available before they start interacting with the system, thus the system will suffer the cold-start problem.

This chapter is organized as follows: in section 3.1, the general model of recommender system developed for this thesis is depicted and some decisions taken affecting the design of the recommender system are presented; in section 3.2, the user profile and how it is constructed are described; in section 3.3, the discussion profile and how it is constructed are described; in section 3.4, the recommendation method used is presented.

3.1. General model

The recommender system developed for this thesis is classified as a hybrid recommender system, unifying the common collaborative filtering method based on users ratings with a content-based approach matching information in both the users' and discussions' profiles.



Conceptual overview of the recommendation scheme developed in this thesis.

As suggested in [Herlocker *et al.* 2004], there are some features which have to be considered when building a recommender system, depending mainly in three factors: domain features, inherent features and sample features.

3.1.1. Domain features

These features represent the general nature of the kind of content being recommended rather than that of the specific system:

- The **items being recommended are online text-based discussions**. The domain where the recommender system will be applied is unknown, as one of its design goals is to be flexible to fit multiple scenarios and multiple domains.
- The user tasks being supported are ***finding good items*** and ***just browsing***, see [Herlocker *et al.* 2004] for the full list of user tasks or use cases in recommender systems (they are summarized in the motivation of this thesis, see section 1.1).

- In a discussion-based environment, the **trade-off between novelty and quality is usually balanced**, although users will consider pointless to be recommended discussions which have been already read unless a fair amount of new comments have been posted to justify the recommendation.
- The **cost/benefit ratio of recommending true/false positives/negatives discussions acquires relevance accordingly to the length of the discussion**. Although is not a critical factor in this scenario, users are supposed to not be willing to waste time reading non-relevant discussions.
- The **granularity of user preferences in a discussion-based environment is assumed to be binary**, *i.e.* "*I like it / I don't like it*", "*that's a discussion I'd like to take part in / I'm not interested in taking part*".

3.1.2. Inherent features

These features represent specific factors introduced by the recommender system from which data is collected, and possibly from its data collection practices:

- **Ratings are explicitly expressed by users**, although other user-preferences heuristics are collected implicitly by the system.
- These **ratings are represented by binary values** having positive values ("1") for expressing positive preference and negative values ("-1") for expressing negative preference. No rating is assumed to show a "certain" degree of disinterest.
- **Ratings are unidimensional**.
- **Ratings data set have time-stamps**.

3.1.3. Sample features

These features reflect distribution properties of the data, and often can be manipulated by selecting the appropriate subset of a larger data set:

- Recommendations displayed to users are not recorded.
- There is no demographic data available for users.

3.2. User profile

The main features which build a user profile in InnoJam are the level of interest in each of the topics, the ratings the user gives to discussions (in fact, this is a shared feature with the discussions users have rated) and the linguistic model the user generates when taking part in the discussions.

3.2.1. Level of interest in topics

The level of interest of the user to each of the topics is represented by a combination of different features, which are implicitly gathered and updated by the system during the interaction with the user:

- The number of discussions the user has viewed for each of the topics being discussed: $viewedDiscussions_i$.
- The number of comments the user has posted to each of the discussions, grouped by topic: $postedComments_i$.
- The number of ratings the user has given to discussions regarding each of the topics, independently if the ratings are positive or negative: $ratingsGiven_i$.
- From those factors, a relative level of interest can be computed for each of the topics with respect to the whole set of topics J, as:

$$relativeInterest_i = \frac{\frac{viewedDiscussions_i}{\sum_{j \in J} viewedDiscussions_j} + \frac{postedComments_i}{\sum_{j \in J} postedComments_j} + \frac{ratingsGiven_i}{\sum_{j \in J} ratingsGiven_j}}{3}$$

3.2.2. Linguistic model

The contributions -comments- a user has made in the system are collected and preprocessed to conform a linguistic model of the user based in a weighted word approach typically used in Information Retrieval for modeling collections of documents. To

The feature is represented in a vector of weighted words where the weight of each of the terms, w_i is computed by the *tf-idf* (term frequency / inter-document frequency) [Salton 1989] measure: $w_i = (tf - idf)_{i,j} = tf_{i,j} \times idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|C_u|}{1 + |\{c : t_i \in c\}|}$, where u represents the user, $n_{i,j}$ is the count of times that term i appears in the comment j, $\sum_k n_{k,j}$ is the sum of occurrences of all other terms in the comment j, which is posted by user u. $|C_u|$ is the total number of comments written by user u, and $|\{c : t_i \in c\}|$ is the number of comments from C_u which contain term t_i.

The vector of weighted words is updated and recomputed after the users write new comments.

3.2.3. Ratings

Users give ratings to discussions when they like or dislike that given discussion. Rating values are then represented by the binary value of the perception by the users, *i.e.* “1” for positive feedback and “-1” for negative feedback.

A new rating is created for a pair of $\langle user, discussion \rangle$ each time a user rates a discussion. The rating is thus represented by a tuple $\langle userId, discussionId, rating, timestamp \rangle$.

In this case, the profile of the user incrementally adds each rating the user performs.

3.3. Discussion profile

The discussion profile models the discussion in terms of its linguistic model and in terms of ratings. Moreover it also has to know which topic the discussion belongs to.

3.3.1. Topic

The discussion profile stores the identifier of the topic that categorizes the discussion. The topic constitutes an important feature to match the discussion profile against the users’ interests.

This feature is considered static, as it is not considered the possibility to change a discussion’s topic.

3.3.2. Linguistic model

As in the user profile, all the comments which conform a discussion are modeled following classic Information Retrieval techniques: after a preprocessing for reducing the term space, consisting in tokenization, stop-word filtering, stemming and a high- and low-frequency terms filtering; a *tf-idf* weighting scheme is applied to the terms as seen in the user profile section.

3.3.3. Ratings

Also as seen in the previous section, the ratings users give to the discussion are part of the profile.

3.4. Similarity measures

Items need to be compared by means of computing the similarity between them. As seen on chapter 2, there are several measures which can be used to compute the similarity between users and between users and discussions.

For computing similarities amongst the levels of interest in topics, Pearson correlation is used. For computing the similarities based on ratings, the Pearson correlation coefficient is used too. For computing similarities between the user and item's linguistic profile, the *tf-idf* term-weighted vectors are compared using the cosine similarity.

3.5. Recommendation method

The recommendation method developed in this work consists on hybridizing two of the most common approaches to the recommendation task, namely collaborative filtering and content-based filtering.

Given the available information and the knowledge we can collect with the system, both collaborative filtering and content-based filtering are useful candidates. No demographic or other kind of external knowledge is provided about the users, so neither demographic filtering nor knowledge-based are not a suitable option. Context-based filtering could be used if taking into account tags, but this approach is considered as future work (see section 7.2.2).

As seen in chapter 2, when hybridizing two or more recommender systems, there are different approaches available. In our case, InnoJam's recommender system uses the switching approach to hybridization. With this approach, the recommender system will use either the collaborative method or the content-based method depending on the available data to make the recommendation.

4. Platform Development

4.1. Concepts

In our approach to innovation jams or innovation forums, which is mainly founded upon the generation and development of ideas through participants' discussions and collaboration, we have based the InnoJam prototype on an existing open source online discussion forum system.

By taking this decision, we rely on a robust and widely tested solution that also provides several security updates and upgrades during the year based on the work of a large group of independent developers.

Before going into the system development itself, we should define which the relevant concepts inside InnoJam are and how they are embodied within the discussion-based system. Although the precise names of these elements may vary a little bit from different forum software systems, they are in general, quite common and recognizable for people with a minimum contact with this kind of system.

4.1.1. Topics

The kind of events that InnoJam aims to give support to, and that Innovation Jams have supported through its different implementations, are based on several topics which participants can discuss about, give their opinions on and provide their knowledge. There is usually a main or general topic -i.e. what the jam is about- but it is discussed through several less general variations of this topic.

The matching element that embodies the topic is the category. A category can be seen as a collection of discussions that belong to or consider the same topic.

4.1.2. Discussions

Discussions are the central element of any forum system. In an innovation forum, these discussions can represent both ideas users can discuss, develop or elaborate and proposals where users can vote on several alternatives.

Discussions can be grouped together thematically into categories or topics. Usually a discussion may only belong to a single category.

A discussion is embodied in a forum system by a thread, which is a composition of a sentence which represents the name or the title of the discussion and a group of comments or contributions.

4.1.3. Comments

Comments are the actual contributions of the participants in terms of textual production -participants can contribute by other means to the jam, e.g. voting, tagging, rating-. Through their comments, participants can discuss, develop and elaborate with other participants.

Comments do conform to an ordered series of collaborative thought. Comments have a context; they are part of a discussion so they may have preceding and following comments.

4.1.4. Proposals

Another kind of contribution that participants can bring into stage is proposals. These proposals can contain some alternatives which are subject for other participants to vote on. Proposals are to be used to gather explicit and discrete information in a specific question.

In terms of online forum systems, proposals and alternatives can be implemented as polls, where participants are given the opportunity to vote on one of the proposed answers to a given question.

4.1.5. Individuals

The participants of the jam, as mentioned early in this document, are in combination with the topics of the event, the real fuel for the jam to produce valuable results. They provide opinion and knowledge through comments and ratings, so building a community-driven participation system.

As seen in chapter 1, crowd-sourcing is what really produces value out of people coming together to discuss several topics from different backgrounds and points of view.

4.2. Specification

4.2.1. Purpose

As discussed previously, in this document, InnoJam provides a solution for fostering the innovation capacity within an innovation process conducted in a company or organization, making it easy to gather the knowledge and creativity of all the actors involved in such a process being part of the mentioned company or organization or not. The solution is based on a discussion platform, where all participants can have a voice and are provided with contextual information that eases their participation and their knowledge about the event.

4.2.2. Scope

Our approach to Innovation Jams relies on the community of participants contributions, thus, the outcome of the jam in terms of ideas and proposals will be conformed out of what was contributed by the participants based on some inputs provided by innovation facilitators, like background information on the topics being discussed.

4.2.3. Functional requirements

Authentication of users

All participants should be authenticated. Authentication should allow access to privileged users to actions related to the administrator or the participant roles.

Define and manage system roles

At least two roles should be taken into consideration for an event like Innovation Jams to function properly.

Some kind of moderator / administrator / facilitator should be in charge of creating the topics that are to be discussed about in the jam, and promoting some of the most relevant or interesting community contributions to a prominent location in the system, like the front page.

Regular users / participants / innovators should be allowed to get into the system and contribute by several means to the generation, development and evaluation of ideas the community undertakes.

Topic based forum

The forum system should allow InnoJam to enable categorized discussions, that is, discussions can belong to only one of a set of categories. Usually this set of categories or general topics to be discussed about should have a size smaller than 10 different categories.

For improved categorization awareness, topics should have distinctive traits, such as color, associated with them.

Create new topics

Administrator / moderator roles should be able to create new categories for participants to discuss about.

Feature participants comments

Administrator / moderator roles should be able to feature relevant or high-quality comments to a prominent space in the system. This space should be, preferably, in the front page of the system.

Allow for participants to discuss among them

The basic feature of InnoJam should be to allow user to participate and contribute. This can basically be achieved by two means: starting new discussions and posting new comments.

Allow participants to make rich multimedia contributions

InnoJam participants should be allowed to base or complement their contributions with rich content elements like external links to other relevant websites or multimedia embeds like YouTube videos.

Allow participants to make proposals

InnoJam should allow participants to make proposals and rely on the community to vote on the several alternatives proposed.

Allow participants to rate contributions

InnoJam participants should be allowed to rate other participants contributions. Contributions that can be rated are discussions and comments. Contributions should be ranked based on the amount of ratings performed by the community.

Allow participants to tag contributions

Content produced by InnoJam participants should be tagged. After contributions have been tagged, several tag-clouds can be produced: global, per category, per discussion and per user tag-clouds.

InnoJam contents should be searchable

Participants should be able to search among the content contributed by other participants in order to find relevant information for them.

Provide participants with contextual information

Participants in the event should be provided with several channels of contextual information, in order to keep users aware of what's going on in the event. Information provided should refer to activity in the discussions, people involved in the discussions, geo-located activity, applied tags to contributions, level of engagement / competence of participants, relations between concepts and / or participants.

Contextual information can be provided by means of visualizations or textual elements.

Provide participants with subscription services

InnoJam should support by some means the diffusion of ideas among the community without creating spam. Some of the proposed alternatives in this area are facilities for sending information to friends or contacts within participant’s social network, feed subscription and viral diffusion.

Generate user profiles for recommendation

The system should generate user profiles, apart from the information provided by the participants that should enable them to get recommendations of either new discussions to participate in or new people to get acquainted with, or both.

4.2.4. Non-functional requirements

For the categorization of non-functional requirements we follow the FURPS+ model used by the Unified Software Development Process [Jacobson *et al.* 1999]. The FURPS+ model makes a distinction between non-functional requirements that are quality requirements and those that are constraints. For quality requirements, we have performance, portability, extensibility, adaptability and flexibility. For constraints we have interface requirements and implementation.

Performance

Although InnoJam is a web application that relies on Internet connection, the user should not experience any performance issue while using InnoJam.

Portability

InnoJam should be easily deployed on standard web servers, without requiring proprietary software.

Extensibility

InnoJam should allow to be extended easily to provide extra functionalities, like more visualizations or more data or text mining techniques.

Adaptability

InnoJam should be easily adapted to work with more or different data parameters or other specific requirements from organizations.

Flexibility

InnoJam should be easily configurable and should allow administrators and innovators to use the system according to their purpose, thus providing several options when adding new topics or making contributions.

Interface requirements

InnoJam should be designed and implemented to provide an easy-to-use environment. As we will see in chapter 5, some efforts will be devoted as well as to easily match projects tools' *look & feel*.

Implementation

Whenever possible, InnoJam should be implemented with open source software and also allow users to access and use it with a regular web browsing environment supporting cookies, flash and javascript.

That implementation should not suppose any inconvenience to most of the potential users of the system.

4.3. Design

One of its main design goals is to provide flexibility to the organizer of the Innovation Jam sessions, meaning that the system supports events discussing either one single topic or several - although it is not recommended to deal with a large amount of topics at the same time, as it can cause problems to the users for following more than 10-15 topics. The user-base can be spread worldwide or very locally.

Another design goal that is implied in the system is the link between colors and topics, i.e. each topic is assigned a different color to represent it. This color-code helps identifying easily all the topics inside the system.

The system is presented in a set of webpages regarding different aspects:

- **Presentation**. It portrays the general status of the Jam session.
- **Discussions**. It lists the ongoing discussions.
- **Comments**. It lists the comments contained in a given discussion. This page is not directly accessible.
- **Topics**. It lists the set of topics proposed to be discussed.
- **Search**. It presents a simple or advanced search form to run searches against the system content.
- **Account**. It contains profile information for the participants in the session.
- **Help**. It contains a contact form and a list of screencasts about InnoJam's usage.
- **Dashboard**. It contains graphical and statistical information relative to the activity in the Jam.

All the pages share the same header, containing the logos of the system and the links to the different pages. Also when the current user does not have an active session in the system, a login form is presented, as well as the traditional links for registering and retrieving a forgotten password.

4.4. Implementation

The core of InnoJam is the open source webforum software Vanilla[23]. Vanilla is a mature forum platform with a stable version (1.x, the one that InnoJam uses) and a development one (Vanilla 2, in beta). The community around it is conformed by developers and designers that build and improve the system by the use of plugins providing new functionalities or look-and-feels, at the same time, most of them are also hosting their own forums, so they know the field first hand.

The main advantages of this package are its robustness, flexibility and its high level of customization, both in terms of function and appearance. Also, coming from the open source world, bug fixes and security upgrades are released fast and often. It has had over 300k downloads and more than 450 plugins, and some of its most relevant clients are O'Reilly Media, Rackspace, Mozilla or Delicious.

We have also made our own **contribution to this open-source community by releasing a couple of our own-developed plugins**, with more than 6k overall downloads.

The full list of plugins running in the system is:

- **PageManager**. This plugin allows the administrator of the system to create new pages in the system, which can be either static -with html- or dynamic -with php-. In InnoJam, it is used to manage the '*Presentation*' page and the '*Dashboard*' page.
- **FirstVisitRedirection**. This plugin allows, in combination with a custom page to redirect the user in her first visit to the system to the '*Presentation*' page.
- **SetList**. This plugin allows plugin creators to create an administration graphical interface with ease. In InnoJam, and other extensions available in the plugin repository of Vanilla, it's used to produce the administrator interface -which enables the administrator to tune or customize some of the parameters of the plugin- of our self-developed plugins.
- **ShowParticipants**. This plugin displays for any discussion a list of the users who have taken part in the discussion. It is a very customizable plugin, as it allows to determine the order of the list and its length, what results, in InnoJam to have the top

[23] Website at *http://vanillaforums.org/*

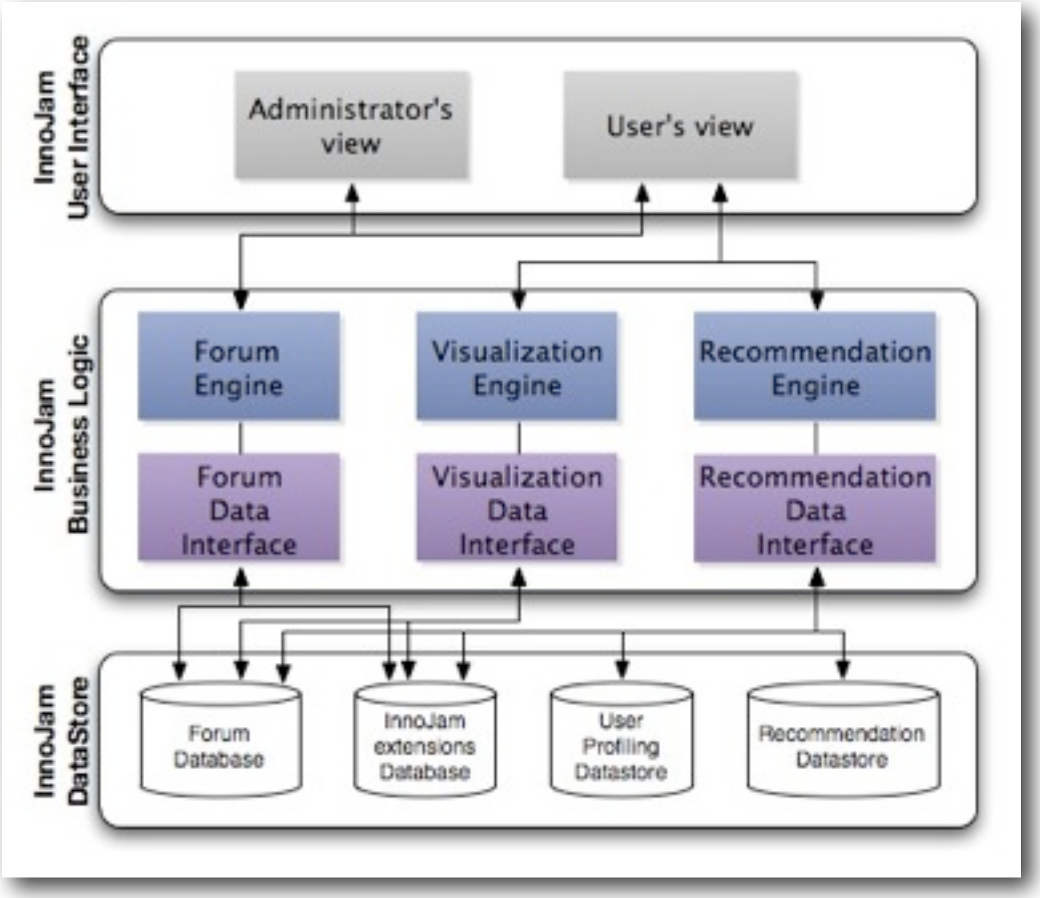
10 commenters in the discussion, and a link to the full list of commenters, knowing beforehand the number of them.

- **PanelTopicsLegend**. This plugin displays in the side panel of several pages a list with the names and colors used in each of the topics.
- **ReportPost**. This plugin is used to allow participants to report inadequate comments to a moderator of the Jam.
- **CommentLinks**. This plugin provides a *permalink* (short for permanent link) for each of the comments, giving then option to other users to share the link with others or for future reference, etcetera.
- **DiscussionExcerpt**. This plugin displays and excerpt of the first comment of the discussion in the discussion list.
- **ParticipatedThreads**. This plugin allows participants to filter the list of discussion to those in which they have posted at least one comment.
- **MakeItSimpleTextFormatter**. This plugin extends the functionality of the commenting input box enabling automatic hyperlinking functions and automatic embedding of multimedia links, like youtube videos.
- **PanelFillerSignIn**. This plugin is a very simple one that just displays a link to sign in for the user if they have an active session -i.e. they are not logged in-.
- **FeedThis**. This plugin provides all the RSS feeds for syndication in the system. The different feeds provided are: discussions feed (all discussions), individual topic feed (all discussions related to a topic), individual discussion feed (all comments in a given discussion), search feed (all comments matching a given search query) and user feed (all comments posted by a given user).
- **GoogleMaps**. This plugin provides all the mashups with Google Maps: the one in the Presentation page for the recent activity and the users, the one for following discussions and also the map on users' profiles where they can set their location.
- **FeaturedComments**. This plugin allows moderators to feature comments to the Presentation page.
- **Tags**. This plugin manages the tags in the comments and produces the tag-clouds in the Presentation page, in the '*Comments*' page, the top tags in the discussion list and the tag-cloud in the user profile.
- **Poll**. This plugin enables users who start a discussion to add a poll to the discussion, thus asking a multi-answer question to other participants.

- **Rating**. This plugin allows discussions and comments to be able to be rated by the participants.
- **UserOrigin**. This plugin allows the system to establish a new field -origin- in the database for each of the users and produce an origin-based version of the statistics and graphics in the '*Dashboard*'.
- **ContactForm**. This plugin allows participants to get in touch and ask support questions, make suggestions, etcetera to the administrator of InnoJam.
- **UserRadarChart**. This plugin generates the visualization of the level of participation per topic in the users' profiles.
- **ProfileComments**. This plugin displays the last comments by user in their profiles.
- **RatingRanking**. This plugin generates the list of discussions sorted by the positive feedback of the community of participants.
- **Announcement**. This plugin allows the administrator of the system to add announcements in several pages of the system, thus serving as a way of communicating with the participants.

The InnoJam functional architecture is composed by three tiers that are loosely coupled:

- **Presentation**. This tier is responsible for the interaction with the user and the system. It presents system views to the user and handles user inputs.
- **Domain**. This tier holds all the business logic of the system. All operations based on user inputs upon data assets are performed in this tier.
- **Persistence**. This tier is responsible for the persistence of the data in the system. It implements storing and loading operations into the data persistence subsystem.



InnoJam 3-layer Architecture

4.5. System Walkthrough

This section provides a shallow view of the different dynamic webpages which conform the system.

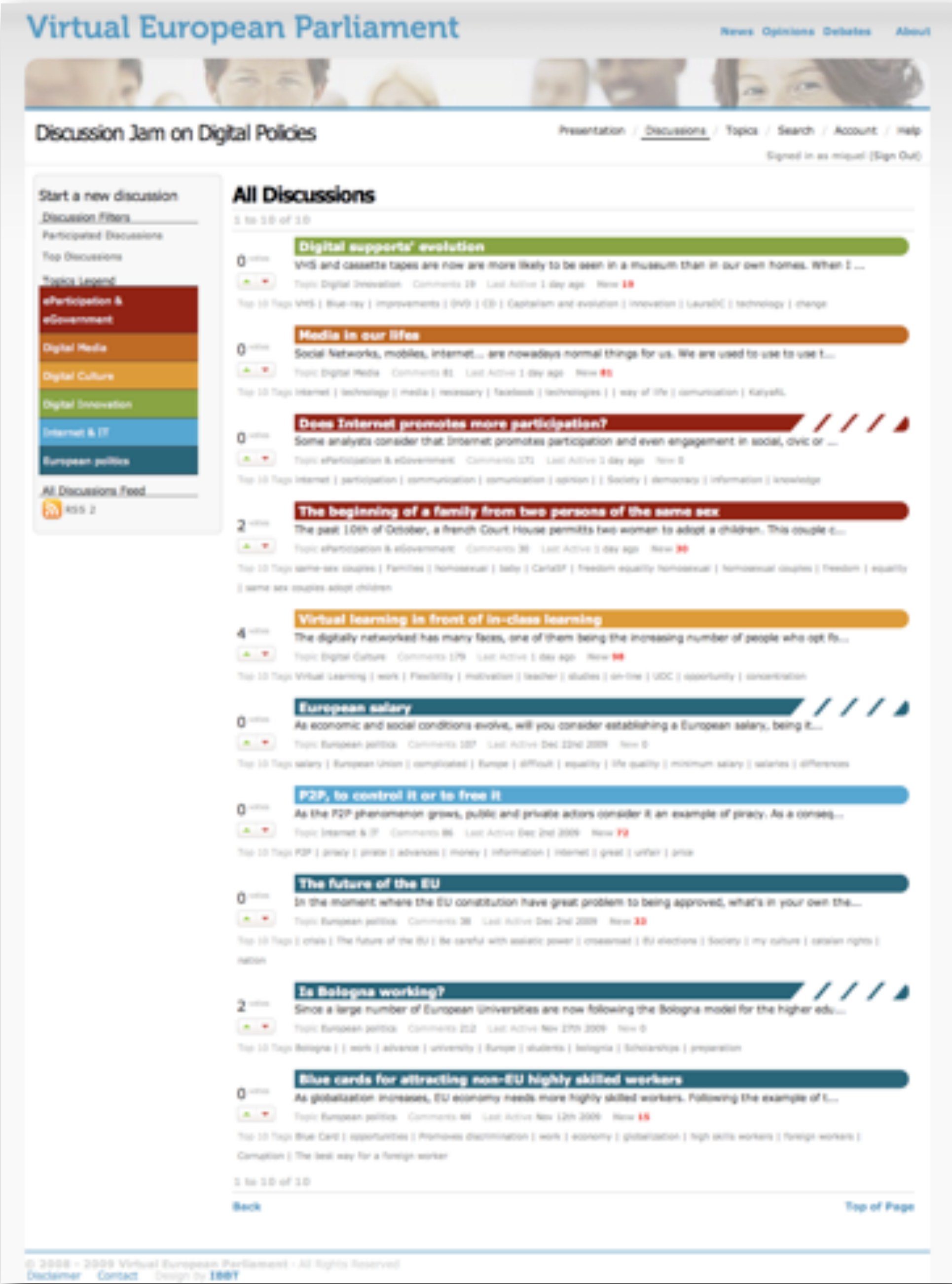
4.5.1. Presentation

Screenshot of the presentation page

This page is intended to be the main entrance point to the system, as it displays information about the session and its activity. The presentation page contains both static and dynamic information about the Jam. Anyway, all the information displayed will be modified according to the system setting up and its working evolution. The pieces of information in the presentation page are:

- The topics menu. A horizontal slider menu containing a visual representation of the proposed topics. Each topic will be represented by a colored shape in the menu, which will resize these shapes depending on the amount of topics discussed. As a minimum size has to be established for the sake of visibility, the solution for displaying a high number of topics in this way is to let the menu move laterally in order to present more topics in the same space. The menu movement is controlled by the user with the mouse cursor. Each topic represented as a colored shape can be clicked and the user will be presented with a list of the discussions related to that topic.
- Geographical activity. A mashup with Google Maps presents both the recent activity in the system and the participants taking part in the Jam which are associated to geographical coordinates, given participants locations. The mashup does automatically open and close sequentially either the comments in a bubble or information about the participants, although the automatic function can be turned on and off depending on viewers' preference. Each of the recent comments is displayed as a colored bubble, according to which topic the comment is tied to.
- Featured interventions. A moderation functionality allows moderators or facilitators to promote featured content to this presentation page. The comment is displayed in the same color as the topic it is related to. Clicking on it will show the complete discussion where the comment was first introduced. Clicking on the comment author's username will show their profile.
- Relevant concepts. Either a treemap -hierarchical visualization- of discussions recently active or a tag cloud with the tags added in the recent comments, or both. The visualization also has the color binding with the correspondent topics.

4.5.2. Discussions



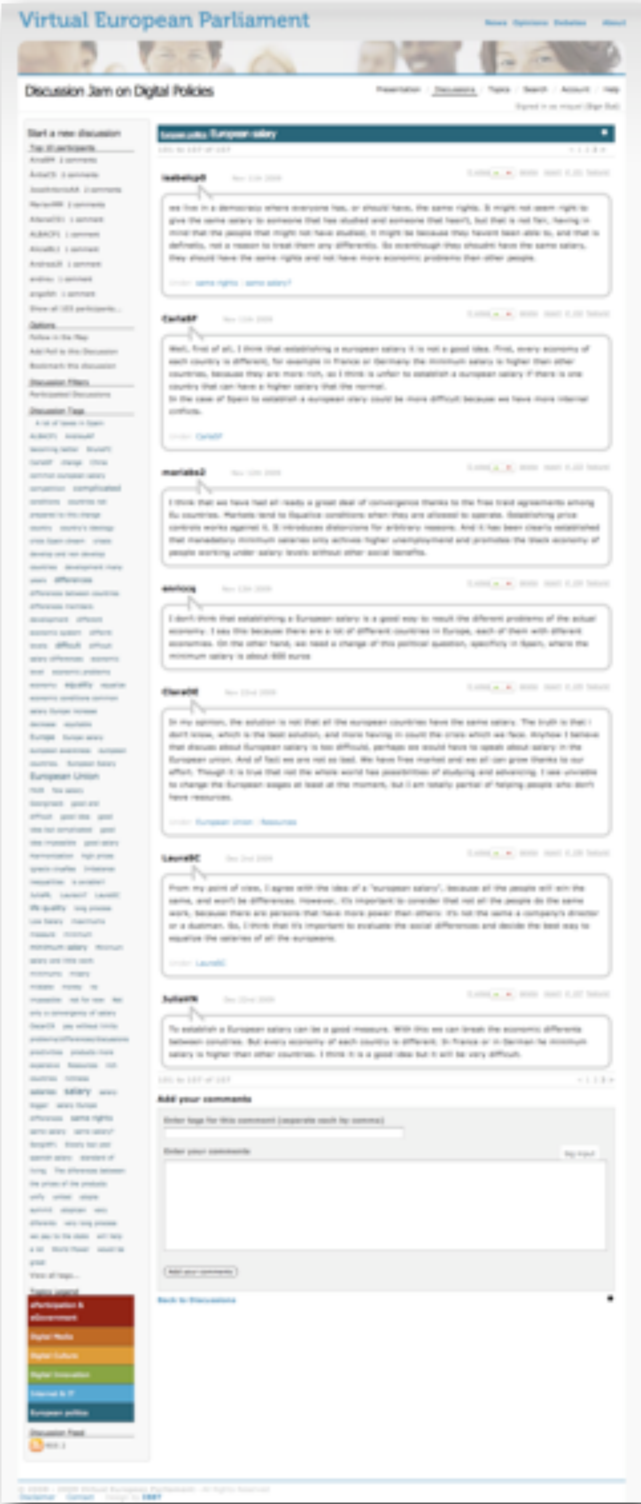
Screenshot of the discussions page

This page contains a list of the discussions where the participants are discussing. This page loads its content dynamically from the database as it serves to several purposes in the system. Its default use corresponds to displaying all the discussions in the system, ordered by freshness, i.e. the most recent they have been updated the upper they get on the list. This view can be accessed from the "*discussions*" link in the header section.

Other views accessible through the discussions page are filtered versions of the default one. These views are, as presented before, the subset of discussions related to a certain topic, the subset of discussions containing a certain tag in one of its comments. Another important list of discussions reorder the original list by the number of positive ratings each of the discussions have received from the participants. In this way, the most praised discussions will naturally float to the top of the list, thus showing their own popularity.

The information displayed in this page concerning discussions consist on the name of the discussion, the topic to which it's related, the user that created the discussion, the number of comments in the discussion, information on who and when the last comment was added and also the amount of unread comments since the user last viewed the discussion -only for registered users-. The information of whether there are new comments in a given discussion since a user's last visit to the system is visually provided too by the color bars in the discussion title. If the bar is fully colored, there are new comments. On the contrary, if there aren't new comments since the last visit, the colored bar is stripped in white.

4.5.3. Comments



Screenshot of the comments page of a discussion

This page contains a list of the comments that belongs to a given discussion. It also offers contextual information about the discussion based on the tags used by the users to tag their comments, the list of participants in the discussion and their respective contributions. It also contains a link for following the discussion on the maps mashup, but in this one, the comments are not automatically showed, but the user is responsible to open them via the selector or following the links inside comment bubbles to previous and next comments (if available). Note, however that the map mashup relies on the users to register themselves with their own geographical location, so only the comments authored by participants who have registered their location on the map via their profile will be displayed.

Screenshot of the map mashup for following a discussion

4.5.4. Topics

Screenshot of the topics page

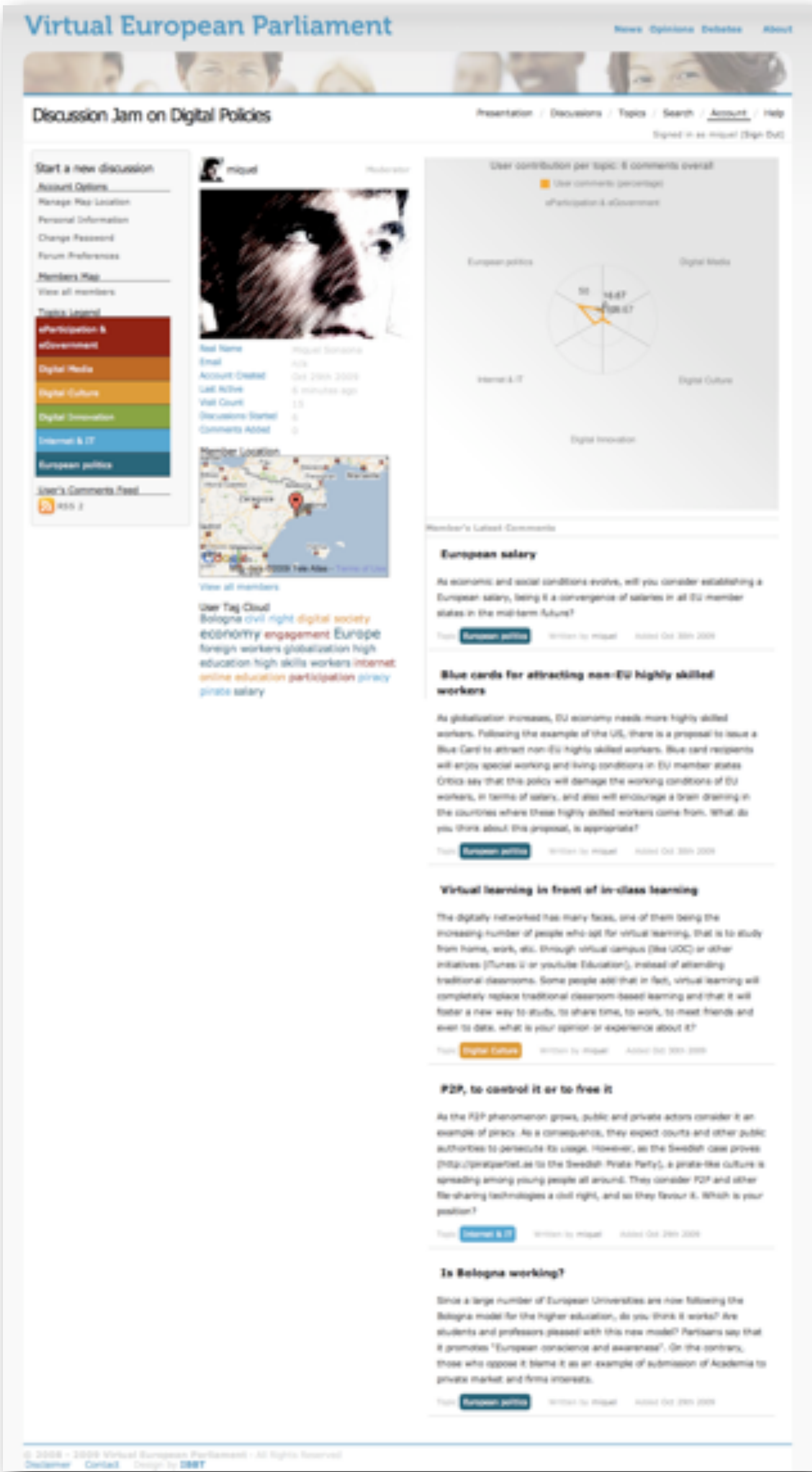
This page contains a list of the topics that participants can discuss about. One key feature is the association between topics and colors, as this association will be present throughout the whole system.

Moreover, each of the topics can include a description and any additional information, like embedded videos or external links to help in framing and giving background information to the discussions in that given topic.

4.5.5. Search

This page contains the form for searching the system content and users. Its basic features are searching a given expression in all the comments, searching for a discussion whose name contains the expression or a user whose username matches the expression entered by the user. The advanced features provide more complex searching methods like defining the author of the comments or discussions the user is looking for.

4.5.6. Account

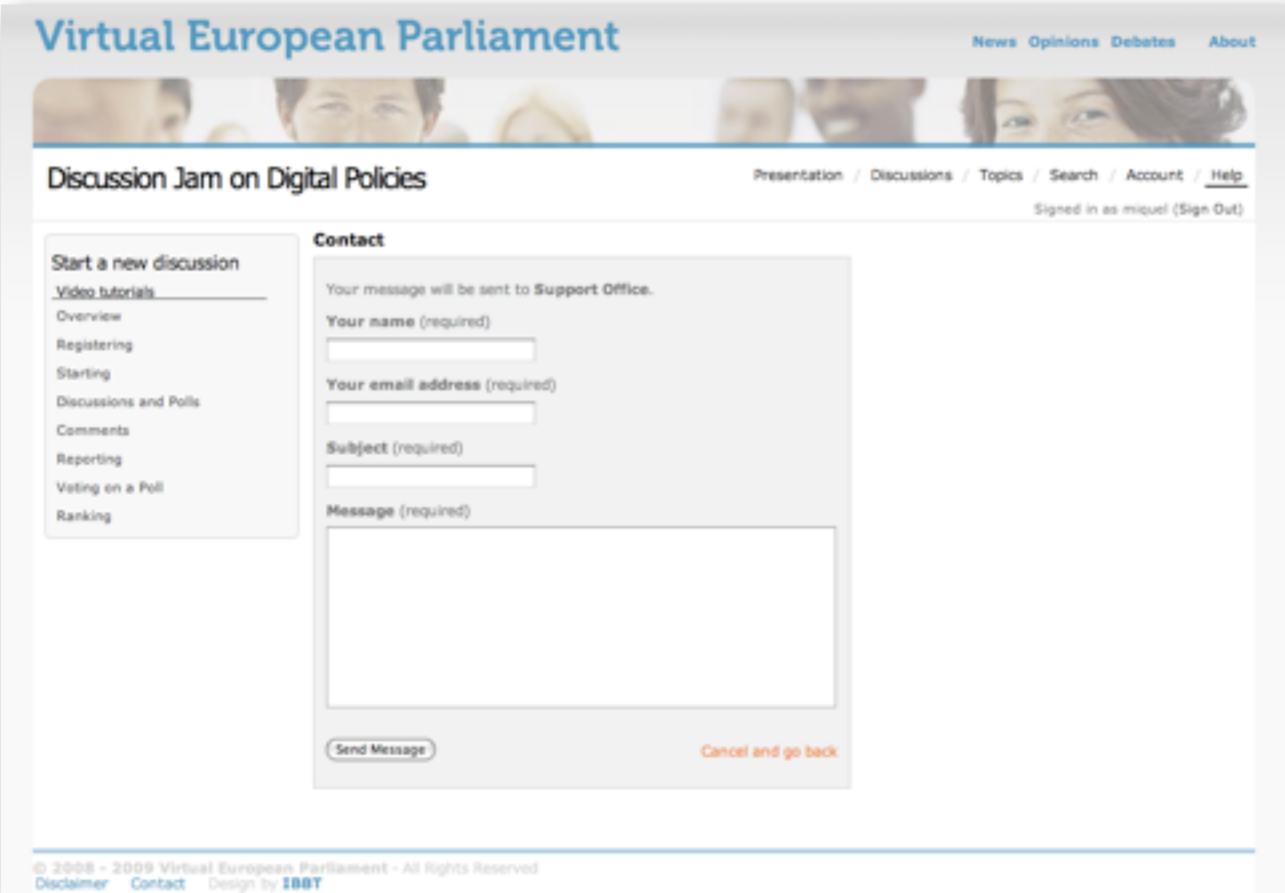


Screenshot of a participant's profile page

The account page represents a user profile, which loads information about a given user. The default status will be that a user accessing this page directly from the header menu link will get his/her own profile. Furthermore, if a user wants to know more about the participant who started a given discussion or posted a comment, they just have to click on the username and the account page will show the information about the discussion-starter or commenter, respectively.

Apart from the standard user-related information that can be found in any user profile, this page also displays the geo-localization of the user -if (s)he sets it up-, the list of latest comments by this user, the tags that have been used by this user and a visualization that depicts the level of participation in each of the topics being discussed in the Jam.

4.5.7. Help



Screenshot of the help page

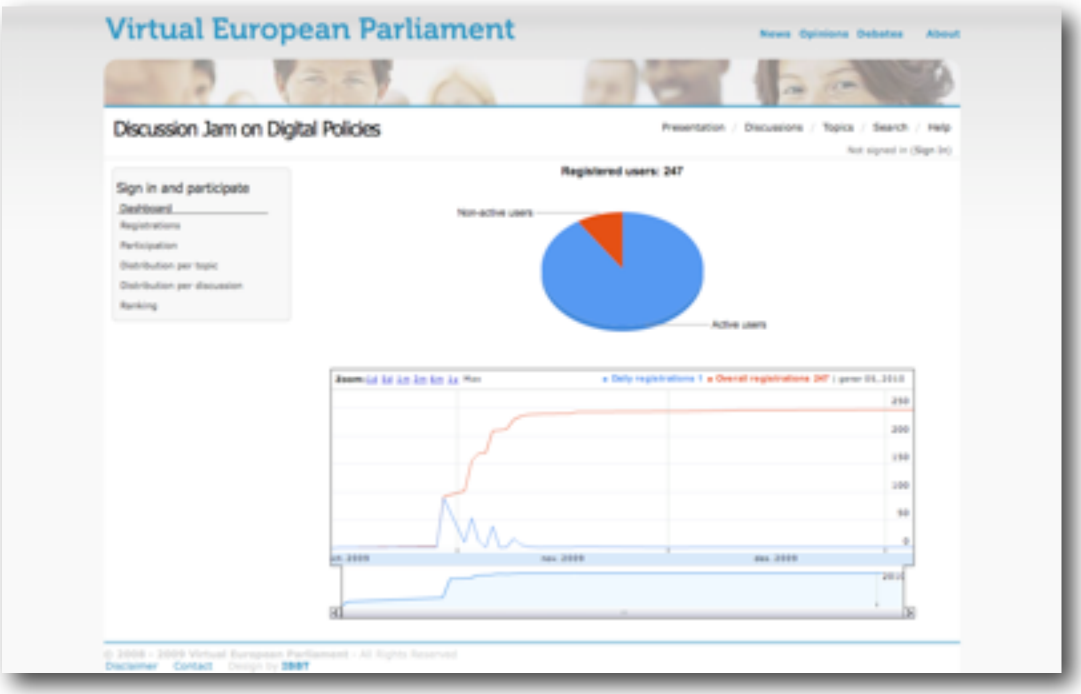
The help page fulfills two purposes, to serve as a contact point between participants and the support office, i.e. the administrators of the event; and to host several video-tutorials about different aspects of the use of InnoJam.

4.5.8. Dashboard

The dashboard contains graphical and statistical information about the level of participation in the Jam. The information displayed is real-time data, so it is a valuable tool to measure the level of activity and engagement of the participants. There are different metrics:

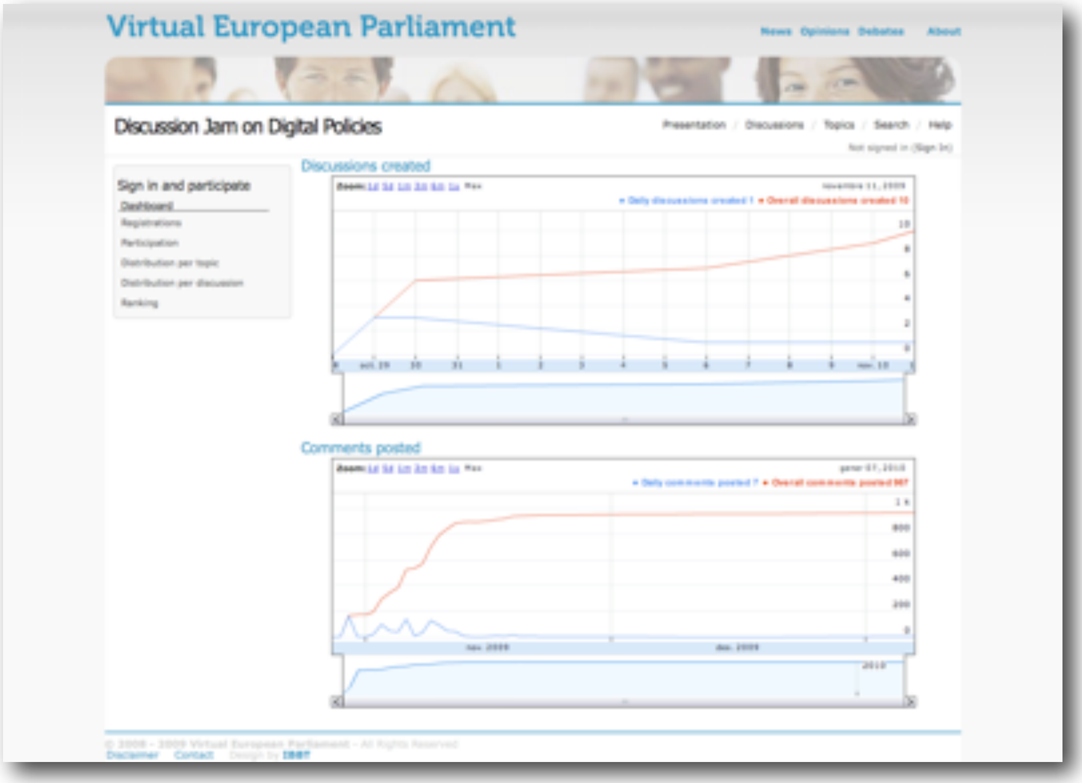
Registration.

It contains information relative to the number and the evolution of registrations over time. It also depicts the percentage of active users.



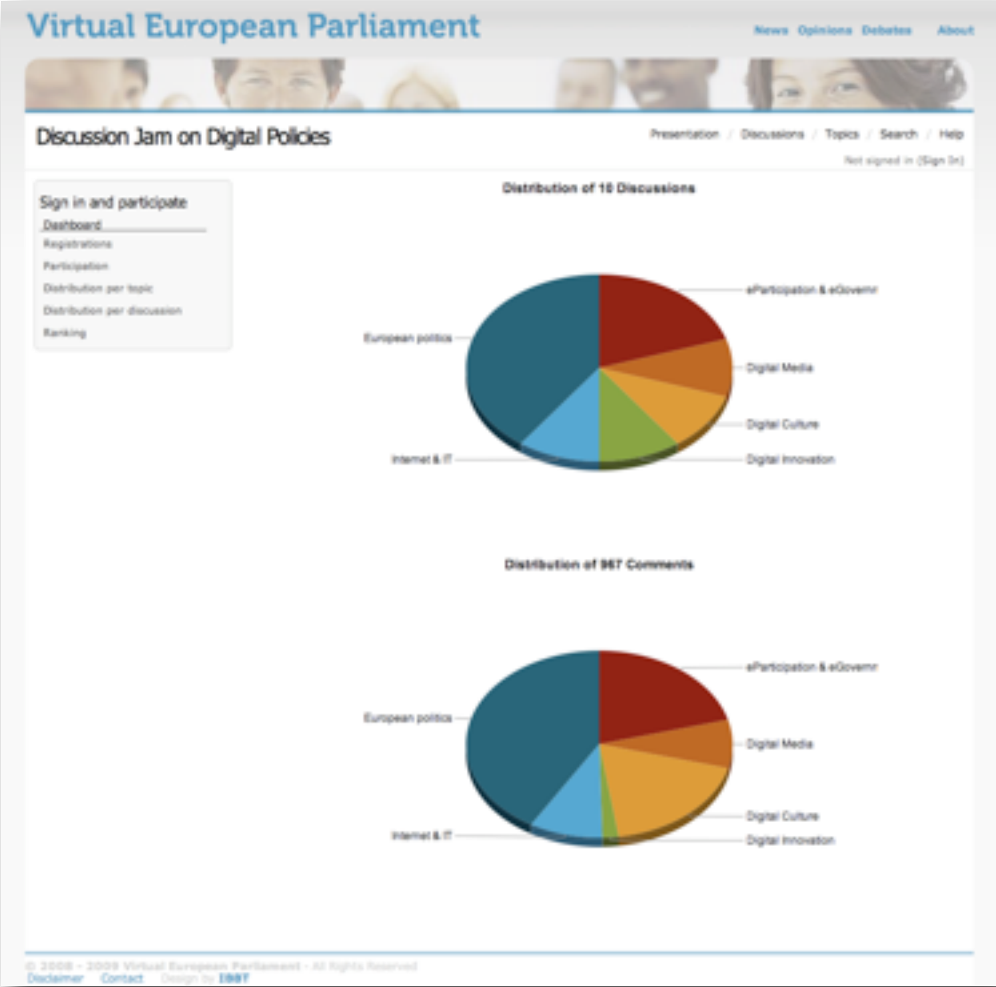
Participation

It contains information relative to the evolution of new discussions and new comments.


Virtual European Parliament
News Opinions Debates About
Discussion Jam on Digital Policies
Presentation / Discussions / Topics / Search / Help
Sign in and participate
Discussions created
Comments posted

Distribution per topic

It contains information of the distribution of discussions and comments per topic.



Distribution per discussion

It contains information on the ratio between the number of comments in the discussion and the number of active participants in that discussion, *i.e.* authors of at least one comment of the discussion.

Ranking

It contains information on the top discussion-creators and top commenters in the Jam.

5. Practical Applications

As mentioned earlier in the document, the development of the thesis has been realized in parallel with, during my internship at the Knowledge Engineering and Machine Learning Group, my contribution to a couple of European Commission funded projects: Virtual European Parliament[24] and Laboranova[25].

As a result of these contributions, we are able to present the practical cases and the environments where InnoJam has been, and could be deployed. Moreover, during the lifetime of this thesis, after meetings with researchers, practitioners and people from diverse backgrounds, and hours and hours of surfing the interwebs, some ideas of alternative uses for InnoJam have arisen.

We will first present accordingly the projects and the role that InnoJam played in them, and afterwards we will present the alternative ideas for its use.

5.1. Virtual European Parliament

Virtual European Parliament, also known as VEP, is a trial European project in the field of eParticipation. The main objective of the project is to bridge the gap between young European citizens and European decision-makers, especially the Members of the European Parliament (MEPs), hence the name of the project. This objective has to be fulfilled by building and deploying a set of Web 2.0 tools, thus building a common room for both citizens and decision-makers to meet. This room is metaphorically embodied in a multi-channel platform, as recommended by eParticipation best practices [Millard 2009].

Nowadays the Internet creates opportunities for youth involvement in politics, and it has some potential to increase young people's political involvement [Kann *et al.* 2007]. An example of this achievement is Mr. Obama's presidential campaign website[26], which

[24] eParticipation call 2007. Grant agreement no. EP-07-01-039. Website at *http://www.virtualep.eu*

[25] 6th Framework Programme. Project no. IST-5-035262-IP. Website at *http://www.laboranova.com/*

[26] Website at *http://www.barackobama.com* (last visit on January 14, 2010)

provided a direct contact between the presidential candidate and his potential voters, thus changing political campaigns from now on [Greengard 2009].As a result of the activity of the project, a multi-channel platform has been built and put into practice.

The multi-channel platform provides young people with the opportunity to be involved in the political discourse by several means:

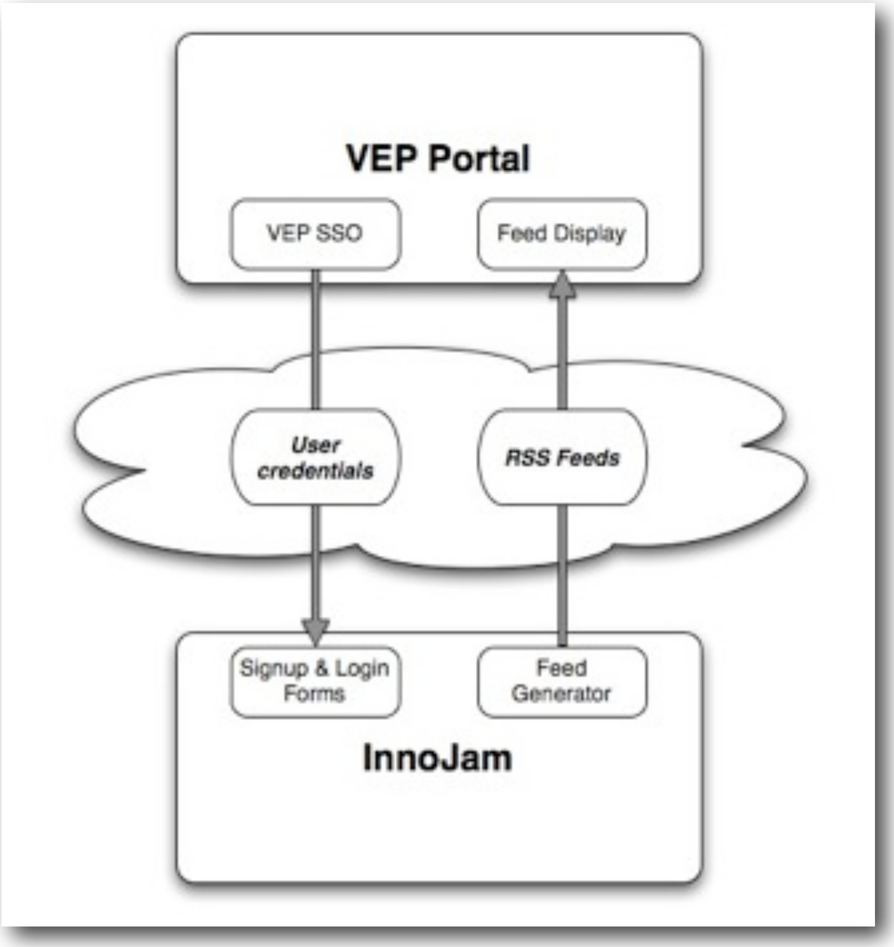
- a Portal, which serves as an entry point to the Virtual European Parliament. It has several integrated functionalities:
  - integrates news feeds from several sources like the European Parliament or blogs related to European issues.
  - host the events pages, which group together information and tools to discuss about a certain topic.
  - integrates blogs, wikis and mobile-voting support.
- MegaConference, a high-definition videoconferencing tool enabling big groups of people to hold plenary meetings, with speaking turns and a streaming service, which broadcasts what's happening in the plenary room and also hosting a chat room and a slideshare for the people connecting from home.
- InnoJam, our tool is used to perform massive debates around certain topics, usually serving as input for a plenary session.
- Flashmeeting, a low-definition videoconferencing tool for communicating a small group of people who are provided by a shared blackboard and online document.

After the first iteration of the project, there were two main concerns in terms of integration of InnoJam with the project portal:

- Common entrance point. Some needs arose from the users in order not to have to register twice (once in the VEP portal, once in InnoJam), so some efforts were taken to overcome this situation and provide a solution which is typically known as a *Single Sign On system*, that is, the users needed just to register once in the portal, and they automatically became registered in InnoJam. The same happened in subsequent visits: users logged in in the portal and automatically became logged into InnoJam. The technical trick was to coordinate several fields in both the portal and InnoJam's databases and to publish the registration and log in forms of InnoJam, *i.e.* the portal which would fill the user credentials in InnoJam automatically.

- Interactivity from the portal. The project needed to centralize somehow the activity in the portal, to focus attention and users from there, so we needed a way to bring content being generated by participants to the portal. The solution was to provide a set of syndication feeds which the portal could display as activity streams. These feeds contained the most recent comments of the Jam session.

A graphical sketch of the integration can be seen in the next figure.



A conceptual scheme of the integration between the project portal and InnoJam

My personal contribution in the VEP project, apart from developing and deploying InnoJam has focused on leading the technical work-package (WP2 - Setup and Maintenance of a Virtual European Platform) and serving as a representative for UPC in the project, thus writing and reviewing deliverables, assisting internal meetings, conferences and review meetings with the EC.

5.2. Laboranova

Laboranova is an European Integrated Project in the 6th Framework Program from the European Commission.

Its full title is Integrated Project Laboranova - Collaboration Environment for Strategic Innovation, and thus aims at creating next-generation collaborative tools for supporting knowledge workers in sharing, improving and evaluating ideas across teams, companies and networks.

Laboranova provides a full-featured toolset which supports the three conceptual pillars the project rely on: ideation, connection and evaluation. The project also provides with facilities for those tools to become integrated together [Oliva and Ceccaroni 2009] and present a common interface and consistent data throughout them.

These facilities include a common look and feel for all the tools, a User repository which will allow users to be authenticated throughout the whole toolset at a time via the Single Sign On component, and the Idea repository which serves as a data store which allows tools to access a central repository where the tools can store their own information and other information which has been produced by their users. The last step in the integration process is the publishing of a set of web services for other tools to consume the information.

My personal contribution in the Laboranova project, apart from developing and deploying InnoJam has focused on collaborating in the so-called "*Connection Space*" (subproject 4) thus writing deliverables, assisting to internal meetings, doing workshops and review meetings with the EC

5.3. Alternative uses

Although InnoJam has been solely applied into the two scenarios previously described, there are others where we consider InnoJam would fit well.

5.3.1. Customer support forum

There are companies which have deployed their customer service in online sites, thus providing a highly accessible way for customer to get in contact with company's representatives. Usually this contact is done via discussion forums where users can deal with different aspects of their relation with the company.

These sites are seldom hosted by the company itself but in specialized web applications like Get Satisfaction[27] or UserVoice[28]. However there are companies that use their own solutions for this purpose like, as mentioned in chapter 1, Starbucks[29] or also Dell[30].

5.3.2. In conferences

InnoJam could be used as a virtual counterpart of conferences, where attendees could discuss with authors and panelists during the whole conference, making these conversations persistent in time.

Given the fact that many attendees at meetings does not really attend, but use the computer in multitasking mode instead, it seems feasible to revert this technological attraction during meetings or conferences into something positive [Benbunan-Fich and Truman 2009].

In this way, the main contributions would be to bring closer to the conference that people that is not attending, thus engaging a much vaster community and to interact more instantly between the attendees [Suter *et al.* 2005].

5.3.3. Question / Answer portal

Question and answer sites, usually known as Q&A are used to publish questions which are online for other (experienced) users to answer them, in order to, gathering individual knowledge, the questions can be answered correctly or at least fulfilling the author expectations. Some famous sites are Yahoo! Answers[31] and, focused on programming issues, Stack Overflow[32].

---

27 Website at *http://getsatisfaction.com/*

28 Website at *http://uservoice.com/*

29 Website at *http://mystarbucksidea.force.com/*

30 Website at *http://www.ideastorm.com/*

31 Website at *http://answers.yahoo.com/*

32 Website at *http://stackoverflow.com/*

6. Evaluation

6.1. Trials with end users

During the VEP project, a couple of trials with young citizens were conducted. The first one was interregional between young citizens from the three member countries. The second was conducted just with young students in Barcelona.

6.1.1. Interregional trial

Having the participation of several dozens of young students among three regions: Luleå (Northern Sweden), Flanders (Belgium) and Catalonia (Spain).

Young citizens were recruited from three different regions and they followed a certain process during the trial where InnoJam took part. They needed a solution for fostering the ideas and opinions of these young students and gathering a set of conclusions in relation with a specific topic, and in this setting, InnoJam fitted well.

The topic which was discussed was '*Blogging*' and among other subtopics, implications with human rights like privacy and security and the quality of blogging were further developed by the participants. For that matter, the project consortium provided some initial content and context in the topic.

In the end, the process was organized as follow:

- In the first phase, the participants, with a little initial prompting by the organizers and some background knowledge acquired from the initial content provided, spent three days of asynchronous discussions between them, getting to know the tool and generating a considerable amount of content by means of starting discussions, introducing comments, applying user-generated tags to either discussions or comments, creating and voting polls and rating other users contributions in the event.
- In the second phase, the participants were divided into two groups, depending on their preference of choice between two subtopics. The groups proceeded with their activity within the system. The selected subtopics were chosen by the organizers of

the event based on their perception of both quantitative and qualitative participation of the young students among all the produced content.

- In the third phase, and based on the outcome of both previous phases, participants were engaged by means of collaborative tools and a videoconference solution to produce a motion of their major conclusions.
- In the fourth and last phase, all participants were called to take part in a final event, following the rules of the European Parliament's plenary sessions. Using a High-Definition videoconference software -named MegaConference- participants were able to address all their peers in the other two countries in real time. Keeping in mind that some participants were from different cities other than the ones which held the event in each country, the MegaConference was also streamed through the web so a subset of the participants could follow the final event from home. Considering that two topics were finally selected, a representative from each country presented their collaboratively generated conclusions in the topic, and the audience could then vote on which of the countries achieved better results with a mobile voting solution, which enabled young students to vote through their cellphones.

As exposed, InnoJam was mainly used as idea generator, aggregator and evaluator in a first stage of the overall process. The outcome of this test can be outlined with some figures.

Overall number of participants	**41 people**
in Belgium	10 people
in Spain	12 people
in Sweden	19 people

Overall number of visits	**333 visits**
in Belgium	65 visits
in Spain	64 visits
in Sweden	204 visits

Overall number of discussions	16 discussions
in Belgium	4 discussions
in Spain	4 discussions
in Sweden	8 discussions

Overall number of comments	133 comments
in Belgium	38 comments
in Spain	30 comments
in Sweden	65 comments

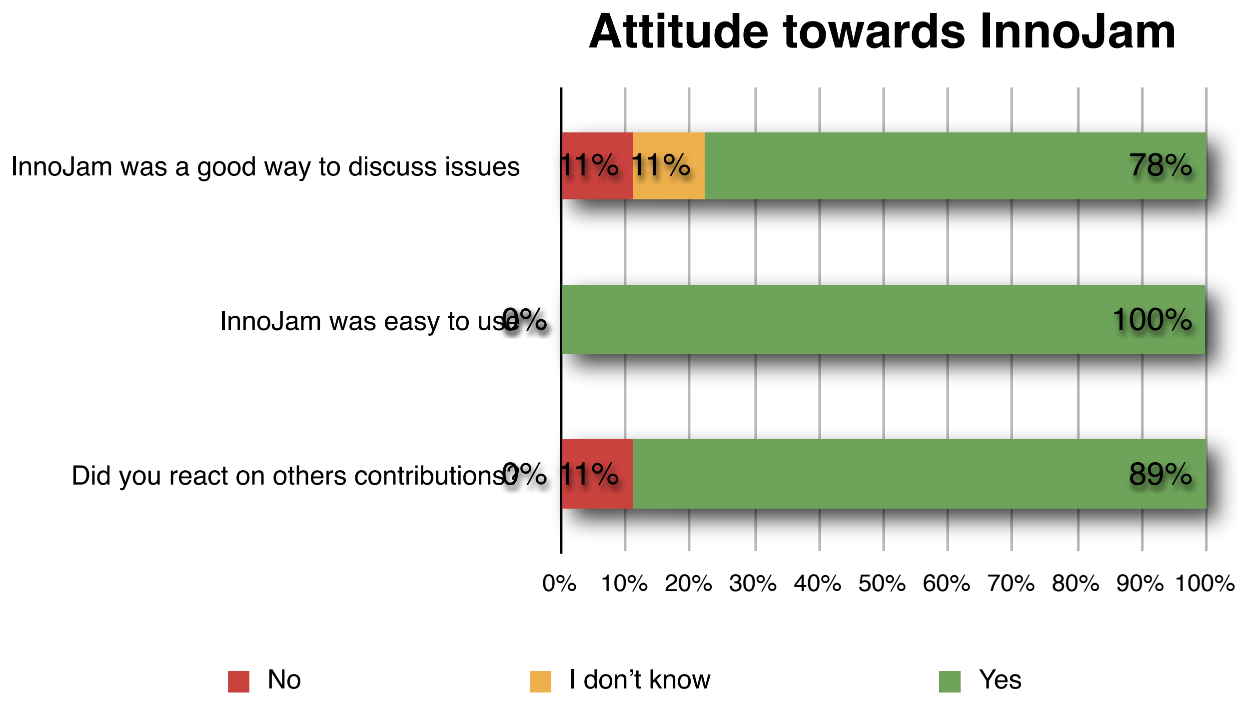
Overall number of applied tags	20 taggings
in Belgium	7 taggings
in Spain	10 taggings
in Sweden	3 taggings

Overall number of ratings	56 casted votes
in Belgium	11 casted votes
in Spain	25 casted votes
in Sweden	20 casted votes

Apart from this quantitative approximation of the usage level of InnoJam by the young participants, they were also prompted to fill in a survey after the whole activities. The participation in this survey is quite low, but we have to point out that some users experienced problems when trying to fill in the survey, so some more users had interest in providing their feedback regarding their experience with the use of the several tools and other concepts like satisfaction and degree of involvement with the activities.

The corresponding part to InnoJam of the survey is as follows:

- General Question: How was it to work with the InnoJam software?

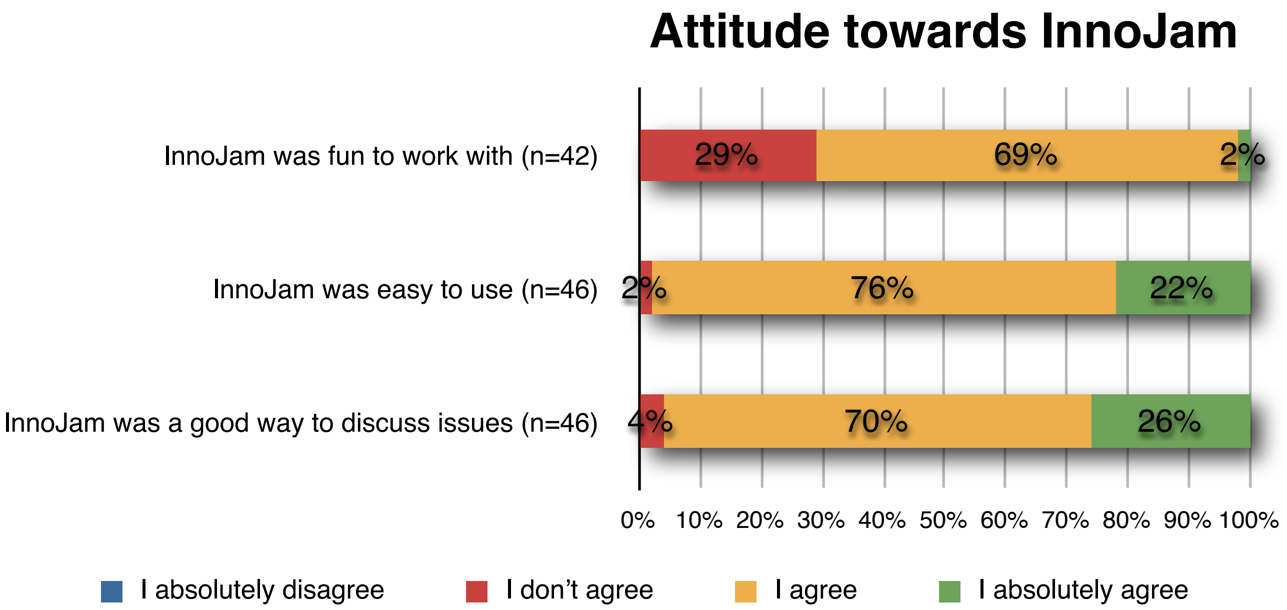


Although the participation in the survey was quite low (9 over 41 participants filled the questionnaire) the percentage of positive feedback is quite high.

6.1.2. Local trial

In this second trial, almost 250 young students from Barcelona were brought together to discuss during a couple of weeks about several proposed topics, all of them related with the intersection between politics and the digital world. The final figures of the event were 10 discussions with more than 950 comments.

After the event finalized, a survey was conducted among participants to assess their experience with the use of InnoJam. The results drawn from the survey were:



6.2. Recommender system evaluation

From these trials with end users, data about their interaction with the system was collected and used afterwards to evaluate the recommender system built for the development of this thesis.

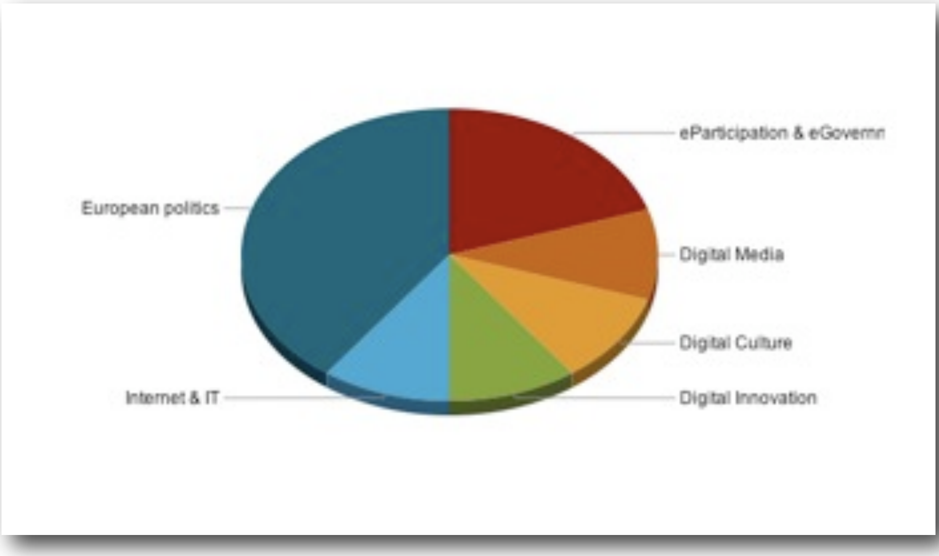
The evaluation has been carried out combining several methods of the recommender system. These components are:

- the collaborative filtering method
  - based in item-item similarity ($I-I \quad CF$)
  - based in user-user similarity ($U-U \quad CF$)
- the collaborative filtering method with the user profile features for computing the similarity between users ($U-U \quad CF+UProf$)
- the content-based method matching the users' profiles to those of the discussions (CB)
- the hybrid recommender combining the last two components ($U-U \quad CF+UProf+CB$)

The validation method used for computing the prediction is *leave-one-out crossvalidation*, where each of the ratings is predicted against the full ratings set as its training data.

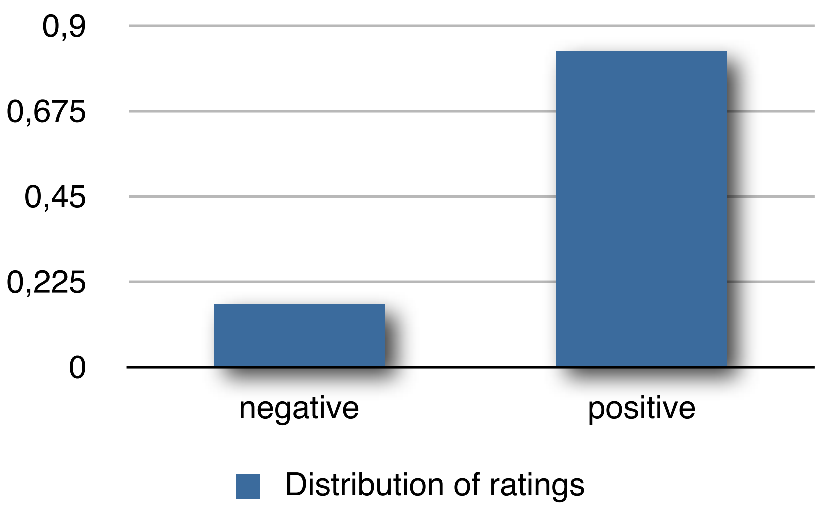
6.2.1. Data set description

The *dataset* is conformed by 247 users and 10 discussions -which represents a ratio of almost 25:1- organized in 6 different topics as follows:

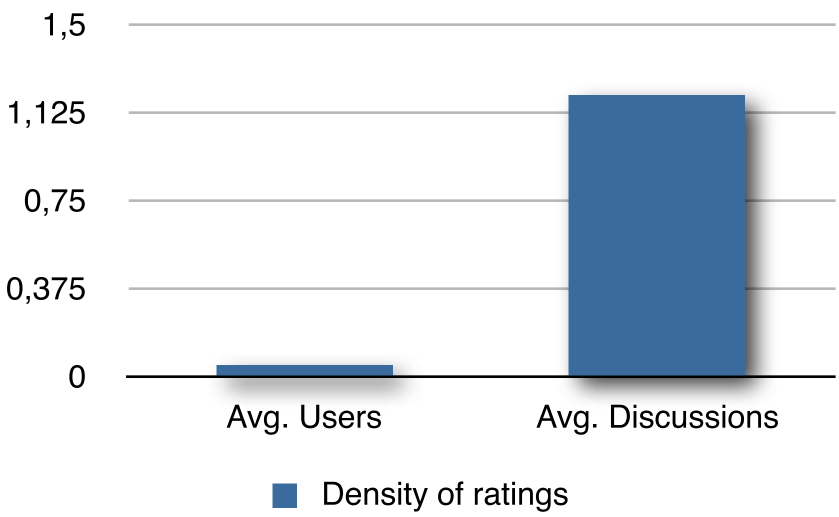


Distribution of 10 discussions among 6 topics

The *dataset* contains 12 ratings given by 10 unique users to 4 different discussions. There is a big bias towards positive ratings:



The density of the ratings set, in terms of the average percentage of discussions that have been rated per user, and the average percentage of users who have rated any discussion.



6.2.2. Metrics

The metric which we have evaluated the recommender system in this work with, the *Mean Average Error,* is one of the most used metric for evaluating recommender systems, especially systems based on the collaborative filtering method. This metric iscomputed with $MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i|$, where p_i is the predicted value and r_i is the real value. The lower the *MAE*, the more accurately the recommendation system predicts user ratings.

6.2.3. Results

Recommendation method	MAE
$I-I \quad CF$	1‘413
$U-U \quad CF$	1‘347
$U-U\ CF + UProf$	1‘168
CB	0‘917
$U-U\ CF + UProf + CB$	0‘764

The previous table reflects the *Mean Average Error* measures for the correspondent configurations of the recommender system.

6.2.4. Discussion

In the evaluation of the InnoJam recommender system, we have focused on measuring the predictive capability of the system by predicting the available ratings based on the rest of the *dataset*.

As could be guessed from the *dataset*, the existence of few ratings has implied that the feature combinations relying solely (or mainly) on the collaborative filtering methods have obtained the poorest results. The absence of a significant number of ratings has a great impact on the collaborative filtering components of the recommender system as they cannot provide a robust and informed prediction.

Otherwise, introducing more informed features about either the users or the discussions carry an improvement on the recommender's accuracy, which obtains recommendations based in other features apart from ratings, so scarce in the tested *dataset*.

7. Conclusions & Future work

7.1. Conclusions

InnoJam is a tool for fostering innovation in massive online events, inspired by IBM's Innovation Jams. This kind of innovation forums have been proven to be successful in gathering knowledge and ideas from innovators' collaboration in a discussion-based manner, thus being a useful tool for companies aiming at unveiling their employees' untapped knowledge or crowd-sourcing this process with large numbers of both internal or external stakeholders.

Our approach, which is built upon an open-source forum software, has a focus on providing activity visualizations and recommendation of discussions. These features will allow participants to become more aware of relevant people and discussions, which will improve the overall outcome of the event.

The objectives of this thesis have been fulfilled:

- The discussion platform is presented and its design and development documented in chapter 4. An open-source forum software has been used as the central core of the system, taking advantage of its robustness and flexibility.
- Also in chapter 4, different visualization mechanisms are depicted. Using several *APIs* mashups and charts have been implemented to render relevant information.
- The hybrid recommender system, combining content-based and collaborative filtering methods, built for InnoJam is presented in chapter 3. The overall system is evaluated from an end-user perspective in chapter 6, which also contains results of the evaluation of the recommender system.

In our trials, users have provided a positive feedback about their use of the tool and its perceived usefulness. These trials served also to gather data for evaluating the hybrid recommender system afterwards.

The evaluation of the recommender system was performed combining the several recommendation methods developed, in order to gain knowledge in which of these methods were more suitable for the task of recommending discussions.

Although the results of the evaluation are not impressive in terms of quality, they depict the fact that the main source of knowledge for collaborative filtering recommenders is the rating matrix, and when it is not big enough or very sparse it produces bad effects on the system's accuracy. On the other hand, introducing other user-profile features or content-based features provides the system with the capability to overcome the problem and obtain better predictability.

Participation should be both qualitative and quantitatively promoted and enhanced in future deployments of InnoJam. More users being more actively involved in creating new discussions, rating and adding tags are needed to take full advantage of the collective intelligence underlying the system. Social science theory suggests reducing the cost of contribution -in terms of design techniques, user goal setting, effort required by the user, etcetera- will increase users' motivation to participate [Beenen *et al.* 2004].

InnoJam could also be used in several other scenarios: getting a company's workforce together to empower innovation, creating a common place for the company members and its customers to innovate in the design of new products or improving the customer services department into a more live and direct relationship, setting up an open forum for municipalities and their inhabitants to arrange common local policies taking into consideration the voice of all the involved actors, and more. InnoJam is flexible enough to support and enhance all these kinds of massive online events.

7.2. Future Work

During the research for this thesis, new ideas have arisen and new research lines have been superficially investigated with potential interest for the continuation of this research work.

7.2.1. Improve the system

One of the clearest lines for future work is iterating on the development of InnoJam, focusing on the improvement of the system and its current set of functionalities. This desired improvement can be outlined in three main areas: the discussion forum core, the visualizations and the integration with social networks.

Discussion Forum

At the time of this writing the open source forum software that powers InnoJam, Vanilla, has its latest version, Vanilla 2, in beta stage[33]. It provides many improvements over the current version 1, which uses InnoJam, like a much powerful underlying framework (named Garden), easier integration with other systems via Single Sign-On[34], improved syndication mechanisms[35] and a better plugin architecture[36] among others. A great novelty of the Vanilla project, which is turning into their business model, is to host the online forum in their own servers instead of a user-contracted one (like Wordpress[37] does with blogs, which can be hosted by Wordpress themselves or the user host the blog in any server they want). This fact could, with some development efforts and publicizing of all our plugins for the community, make possible for InnoJam to be self-hosted online too.

Visualization

Information visualization is a very active field both in research and in industry, specially in news agencies and newspapers infographic offices. In these times, the real-time Internet requires powerful visualization tools to filter and aggregate an overwhelming amount of information both in quantity and in time, and to display trends whenever and wherever they are sparking.

Interactive visualization depicting movie revenues along time[38]

---

[33] Website at *http://vanillaforums.org/*

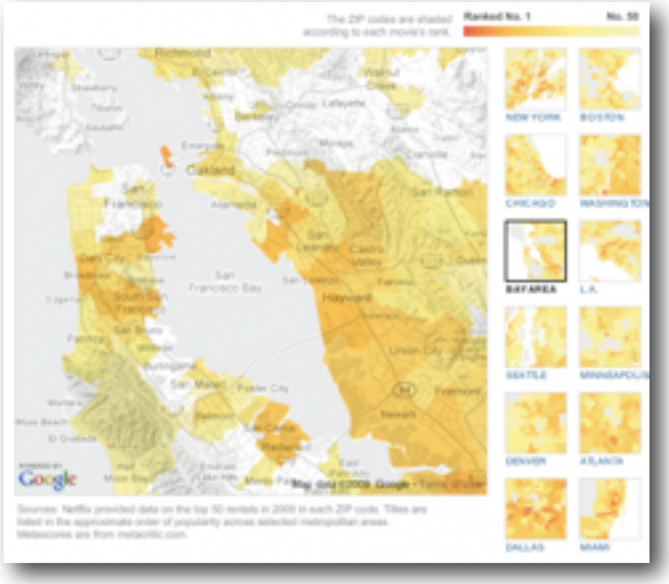
[34] Website at *http://vanillaforums.org/page/SingleSignOn*

[35] Website at *http://vanillaforums.org/page/Syndication*

[36] Website at *http://vanillaforums.org/page/Plugins*

[37] Wordpress online blogs at *http://wordpress.com* and downloadable hosted version at *http://wordpress.org*

[38] Accessible at the NY Times website at *http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html*



Infographic depicting movie renting popularity[39] using a heatmap over Google Maps

Integration with social networks

In the Internet nowadays it makes sense to **integrate web applications with the most common social networks**, like Facebook[40] or twitter[41], in order to allow participants on the Jam to share their thoughts with their friends or followers in these social networks and thus potentially attracting a far wider audience in the Jam.

Another interesting line of research is to investigate how to use these social media channels and platforms in order to improve the level of engagement and the participation of users, like act.ly[42] intends to achieve.

7.2.2. Recommender System

Concerning the recommender system, the main focus for future work should be to **test the recommendation engine in a live environment**, with real users using the system. This would give us a real measure of the goodness of the recommender system, because, in the end, the recommender system is just a tool for serving a purpose, namely helping users to find interesting discussions to take part in; and this can only be assessed by testing it in a real environment.

---

[39] Accessible at the NY Times website at *http://www.nytimes.com/interactive/2010/01/10/nyregion/20100110-netflix-map.html*

[40] Website at *http://www.facebook.com/*

[41] Website at *http://twitter.com/*

[42] Website at *http://act.ly/*

As seen on chapter 2, one of the main factors which affects collaborative filtering recommender systems is the **cold start** problem. Although the system we have designed does not suffer from this problem regarding new items, as the discussion profile is directly initialized when the discussion is created, we need a solution for this problem regarding new users. One of the most viable solutions we can devise currently is to **force the users to explicitly provide their levels of interest for the different topics being discussed when registering** in the system, thus explicitly building their initial profile. E.g. with *sliders* or a design mechanism for re-ranking the topics to reflect more accurately the order of preferences.

Another factor a recommender system has to deal with is **novelty**. A trade-off exists between the novelty of a recommended item and the familiarity of the user with the item being recommended, especially when recommending items for new users. In this scenario, one of the improvements could be to **define a balance factor to this trade-off**. This factor should have a distribution based either on the longevity of the user in the system or on the usage level this user has within the system, thus providing more obvious recommendations to new users and more novel recommendations to long-lasting or more experienced users.

Web Mining

Seen as the discovery of new information by the analysis of the content, the structure and the usage of web systems, web mining may provide us with improvements in the **description of discussions** and their **similarity** and the **habits of users and their preferences**.

Social Tagging

Another useful research line, given the characteristics of our environment, is to make use of a third dimension in the recommendation process a part from the users and items. This third dimension is the use of tags which can be used to enrich both users and discussions profiles and enhance the conceptual association between users and discussions [Celma 2008, Mathes 2004].

7.2.3. Active Learning

Active learning constitutes an interesting research line, which is quite active in recent times [Stemp-Morlock 2009]. The idea behind active learning is to crowdsource a complex learning task to human beings, usually through the use of serious games [von Ahn 2006, von Ahn and Dabbish 2008]. In this sense, one of the possible next steps is to feature an active learning approach into the system, which would consist in getting the user playing a serious game in order to **produce new tags** explaining the contents produced by other participants, as in Luis von Ahn's *the ESP game* [von Ahn and Dabbish 2004] or in Jeffrey Orkin's *The Restaurant Game* [Orkin 2007, Orkin and Roy 2007].

This approach will give us a larger and more representative tagging as multiple users will tag a given comment, hence enriching both quantitatively and qualitatively the current tagging of the discussions. This improvement will, as a consequence, enhance the folksonomy [Mathes 2004] associated with comments and discussions, the representation of comments' concepts and the overall recommendation process, as seen in previous section.

7.2.4. Prediction Market

In prediction markets [Wolfers and Zitzewitz 2004], ideas, or other "products", have a stock price and the process of evaluating a collection of ideas becomes a stock market, where participants buy and sell ideas. This price is the expectation of the users of the future price or value of the idea they are buying or selling. In this way, a prediction market is a system that captures collective knowledge [Watkins 2007].

Adding this prediction market feature will most notably change how the ideas are rated, but will in the end provide a similar outcome in the overall result: the community decides upon which are the most relevant ideas and which are not. However this approach may be useful in a more corporate environment, as it can be seen as **a more formal way of idea evaluation** than just giving thumbs up or down.

7.2.5. Natural Language Processing

Being InnoJam a textual discussion-based system, it makes sense to further investigate the chances Natural Language Processing offers us for improving the system.

Automatic summarization

One of the NLP applications which could enhance InnoJam is automatic summarization, providing us with another useful tool to **reduce the information overload** inherently present when large communities interact.

Machine translation

Another NLP application that could be fruitful in our future research is automatic translation. We have found the need when interconnecting communities from different countries (see chapters 5 and 6), that although using English as their vehicular language, most of them would have benefitted from **using their own language for better expressing themselves** and their points of view.

Even though, this scenario could become quite complex as the system should be able, in the end, to translate the content produced by participants into all the other languages used, an intermediate solution for this problem could be to translate all content from the original language used by the authors into English, and then from English translate it into all the other languages used by all participants. In this way, we have the original language version for participants using the same language and the version translated from English for the rest of participants in their own language, which will result in a presumably better translation than directly from the original language.

Other ideas

Other applications that might have relevance in our scenario, but have just been very shallowly considered are **opinion extraction**, **sentiment analysis** and **trend identification**. These applications could provide a **higher understanding of the ongoing conversations** in the event, which could lead to a more informed recommendation of discussions.

References

Adomavicius, Gediminas; Sankaranarayanan, Ramesh; Sen, Shahana; Tuzhilin, Alexander. 2005. **Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach.** ACM Transactions on Information Systems, volume 23, number 1, January 2005, pp. 103-145.

Adomavicius, Gediminas; Tuzhilin, Alexander. 2005. **Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions.** IEEE Transactions on Knowledge and Data Engineering,volume 17, number 6, June 2005.

Basilico, Justin; Hofmann, Thomas. 2004. **Unifying Collaborative and Content-Based Filtering.** Proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

Beenen, Gerard; Ling, Kimberly; Wang, Xiaoqing; Chang, Klarissa; Frankowski, Dan; Resnick, Paul; Kraut, Robert E. 2004. **Using Social Psychology to Motivate Contributions to Online Communities.** CSCW'04, November 6-10, 2004, Chicago, Illinois, USA.

Benbunan-Fich, Raquel; Truman, Gregory E. 2009. **Multitasking with laptops during meetings.** Communications of the ACM, volume 52, number 2, February 2009.

Bennett, James; Lanning, Stan. 2007. **The Netflix Prize.** ACM KDDCup'07.

Bjelland, Osvald M.; Chapman Wood, Robert. 2008. **An Inside View of IBM's 'Innovation Jam'.** MIT Sloan Management Review, volume 50 number 1, Fall 2008, pp. 31-40.

Burke, Robin D. 2002. **Hybrid recommender systems: survey and experiments.** User Modeling and User-Adapted Interaction, 2002.

Celma, Òscar. 2008. **Music recommendation and discovery in the long tail.** Ph.D. Thesis. Department of Information and Communication Technologies, UPF.

Codina, Victor. 2009. **Design, development and deployment of an intelligent, personalized recommendation system.** Thesis, Master in Artificial Intelligence. UPC.

Gabor, Andrea. 2009. **The Promise (and Perils) of Open Collaboration**. strategy +business, issue 56 (Autumn 2009). Accessible at *http://www.strategy-business.com/article/09302*

Goldberg, David; Nichols, David; Oki, Brian M.; Terry, Douglas. 1992. **Using collaborative filtering to weave an information tapestry.** Communications of the ACM, volume 35 , issue 12 (December 1992).

Greengard, Samuel. 2009. **The First Internet President**. Communications of the ACM, volume 52, number 2 (February 2009), pp. 16-18.

Herlocker, Jonathan L.; Konstan, Joseph A. 2001. **Content-Independent Task-Focused Recommendation.** IEEE Internet Computing, volume 5, number 6, 2001, pp. 40-47.

Herlocker, Jonathan L.; Konstan, Joseph A.; Terveen, Loren G.; Riedl, John T. 2004. **Evaluating Collaborative Filtering Recommender Systems.** ACM Transactions on Information Systems, Vol. 22, No. 1, January 2004, Pages 5–53.

Hoffmann, Leah. 2009. **Crowd Control.** Communications of the ACM, volume 52, number 3 (March 2009), pp. 16-17.

Howe, Jeff. 2006. **The Rise of Crowdsourcing.** Wired, volume 14, number 6, 2006. Accessible at *http://www.wired.com/wired/archive/14.06/crowds.html*

IBM. 2009. **IBM Global CIO Study 2009**. Downloaded from *http://www-935.ibm.com/services/us/cio/ciostudy/*

Jacobson, Ivar; Booch, Grady; Rumbaugh James. 1999. **The Unified Software Development Process.** Addison Wesley, Reading, Massachusetts 1999.

Kann, Mark E.; Berry, Jeff; Gant, Connor; Zager, Phil. 2007. **The Internet and youth political participation**. First Monday, volume 12, number 8 (August 2007).

Kohrs, Arnd; Merialdo, Bernard. 2001. **Improving collaborative filtering for new-users by smart object selection.** Proceedings of International Conference on Media Features (ICMF), 2001.

Krulwich, Bruce. 1997. **LIFESTYLE FINDER: Intelligent User Profiling Using Large-Scale Demographic Data.**

Leino, Juha; Räihä, Kari-Jouko. 2007. **Case Amazon: Ratings and Reviews as Part of Recommendations.** ACM RecSys'07, October 19–20, 2007, Minneapolis, Minnesota, USA.

Linden, Greg; Smith, Brent; York, Jeremy. 2003. **Amazon.com recommendations: Item-to-item collaborative filtering.** IEEE Internet Computing January-February 2003.

Mathes, Adam. 2004. **Folksonomies - Cooperative Classification and Communication Through Shared Metadata.** Computer Mediated Communication - LIS590CMC. Accessible from *http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html*

McDonald, David W. 2003. **Recommending collaboration with social networks: a comparative evaluation.** Proceedings of the SIGCHI conference on Human factors in computing systems 2003, Ft. Lauderdale, Florida, USA.

McNee, Sean M.; Riedl, John; Konstan, Joseph A. 2006. **Accurate is not always good: How Accuracy Metrics have hurt Recommender Systems.** CHI 2006, April 22-27, 2006, Montreal, Canada.

Millard, Jeremy. 2009. **eParticipation Recommendations - focusing on the European level.** Deliverable D5.1 European eParticipation. Accessible at *http://www.european-eparticipation.eu/index.php?option=com_docman&task=cat_view&gid=36&&Itemid=82*

O'Reilly, Tim. 2005. **What is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software.** Accessible at *http://oreilly.com/web2/archive/what-is-web-20.html*

Oliva, Luis; Ceccaroni, Luigi. 2009. **REST Web Services in Collaborative Work Environments.** Proceedings of the 12th International Conference of the Catalan Association for Artificial Intelligence. pp. 419-427.

Orkin, Jeff. 2007. **Learning Plan Networks in Conversational Video Games**. Masters Thesis, MIT Media Lab.

Orkin, Jeff; Roy, Deb. 2007. **The Restaurant Game: Learning Social Behavior and Language from Thousands of Players Online**. Journal of Game Development, 3(1), pp. 39-60.

Procter & Gamble. 2006. **P&G Connect + Develop Open Innovation**. Slideshare accessible at *http://www.scribd.com/doc/14757218/PG-Connect-Develop-Open-Innovation*

Ramezani, Maryam; Bergman, Lawrence; Thompson, Rich; Burke, Robin; Mobasher, Bamshad. 2008. **Selecting and Applying Recommendation Technology.** International Workshop on Recommendation and Collaboration, 2008 International ACM Conference on Intelligent User Interfaces (IUI 2008), Canary Islands, Spain, January 2008.

Resnick, Paul; Iakovou, Neophytos; Sushak, Mitesh; Bergstrom, Peter; Riedl, John. 1994. **GroupLens: an open architecture for collaborative filtering of netnews.** CSCW'94.

Rubens, Neil; Sugiyama, Masashi. 2007. **Influence-based Collaborative Active Learning.** RecSys'07, October 19–20, 2007, Minneapolis, Minnesota, USA.

Salton, Gerald.1989. **Automatic Text Processing.** Addison-Wesley, 1989.

Sarwar, Badrul M.; Karypis, George; Konstan, Joseph A.; Riedl, John T. 2000. **Application of Dimensionality Reduction in Recommender System - A Case Study.** ACM WebKDD 2000

Sarwar, Badrul M.; Karypis, George; Konstan, Joseph A.; Riedl, John T. 2001. **Item-Based Collaborative Filtering Recommendation Algorithms.** WWW10, May 1-5, 2001, Hong Kong.

Shardanand, Uppendra; Maes, Pattie. 1995. **Social Information Filtering: Algorithms for Automating 'Word of Mouth'.** Conference on Human Factors in Computing Systems, 1995.

Sonsona, Miquel; Almirall, Esteve. 2009. **Enhancing Innovation with Social Recommendations.** ICE 2009 Proceedings, June 22-24, 2009, Leiden, The Netherlands.

Spangler, W. Scott; Kreulen, Jeffrey K.; Newswanger, James F. 2006. **Machines in the Conversation: detecting themes and trends in informal communication streams.** IBM Systems Journal, Vol 45, No 4, 2006.

Stemp-Morlock, Graeme. 2009. **Learning More About Active Learning**. Communications of the ACM, volume 52, number 4 (April 2009), pp. 11-13.

Surowiecki, James. 2004. **The Wisdom of Crowds.** Random House, 2004.

Suter, Vicki; Alexander, Bryan; Kaplan, Pascal. 2005. **Social software and the future of conferences right now.** Educause Review, January-February 2005 pp. 46-59.

Tapscott, Don; Williams, Anthony D. 2006. **Wikinomics: how mass collaboration changes everything.** Atlantic Books, London.

Tejeda, Arturo. 2006. **Sistema de recomendación turística basado en RBR y CBR.** Master Thesis, ITESM, Mexico.

Tejeda, Arturo; Sancho, Agustí; Almirall, Esteve. 2009. **Xpertum: Competences and Social Networking.** ICE 2009 Proceedings, June 22-24, 2009, Leiden, The Netherlands.

Terveen, Loren; McDonald, David W. 2005. **Social matching: A framework and research agenda.** ACM Transactions on Computer-Human Interaction (TOCHI), volume 12, issue 3 (Semptember 2005), pp. 401-434.

Tintarev, Nava; Masthoff, Judith. 2007. **Effective Explanations of Recommendations: User-Centered Design.** RecSys’07, October 19-20, 2007, Minneapolis, Minnesota, USA.

von Ahn, Luis. 2006. **Games with a Purpose**. IEEE Computer Magazine, volume 39, issue 6 (June 2006), pp. 92-94.

von Ahn, Luis; Dabbish, Laura. 2004. **Labeling Images with a Computer Game**. CHI 2004, April 24–29, 2004, Vienna, Austria.

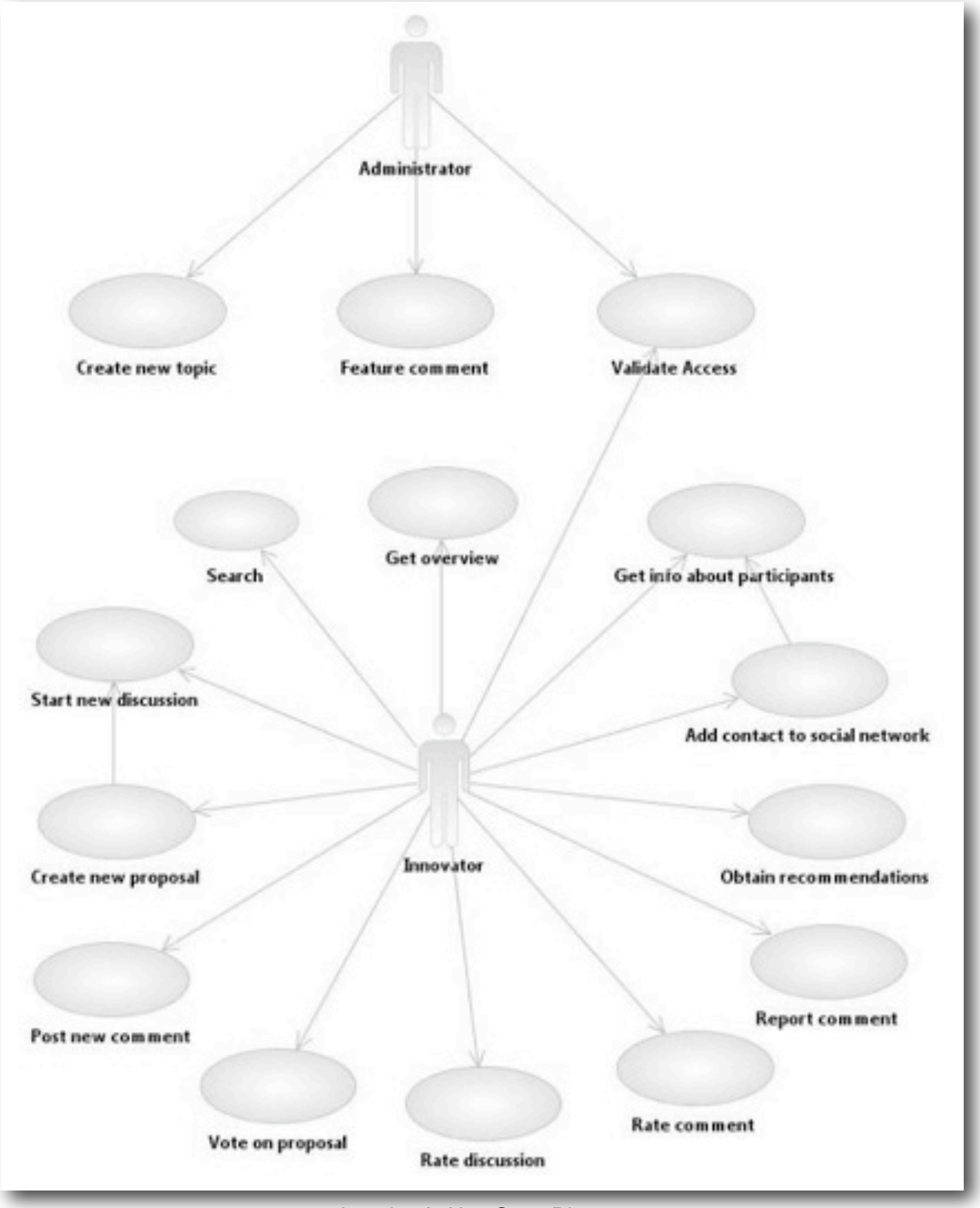
von Ahn, Luis; Dabbish, Laura. 2008. **Designing Games With A Purpose**. Communications of the ACM, volume 51, number 8 (August 2008), pp. 58-67.

Watkins, Jennifer H. 2007. **Prediction Markets as an Aggregation Mechanism for Collective Intelligence.** Proceedings of 2007 UCLA Lake Arrowhead Human Complex Systems Conference, Lake Arrowhead, CA, 25-29 April 2007.

Wolfers, Justin; Zitzewitz, Eric. 2004. **Prediction markets.** Journal of Economic Perspectives, volume 18, number 2 (Spring 2004), pp. 107-126.

Appendix - Use case model

This section describes the use case model of InnoJam, derived from the analysis of the requirements identified in chapter 4.

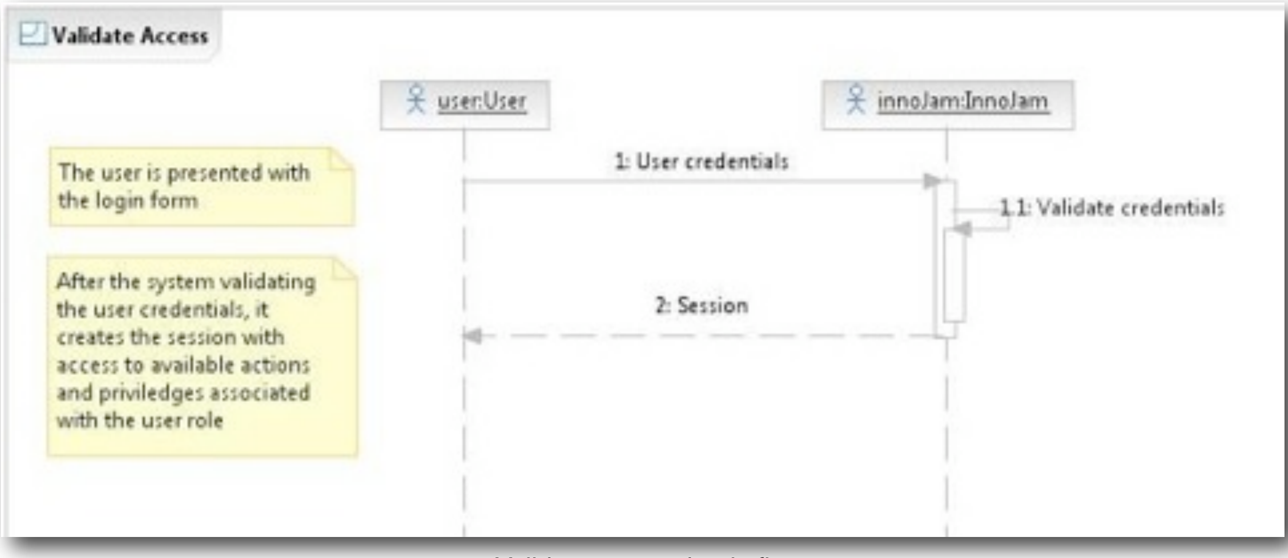


InnoJam's Use Case Diagram

Validate access

Users aim at accessing the system with their user credentials (user-name and password) and the system, upon checking these credentials gives or denies access to the system.

Depending on these credentials, as mentioned before, the user will be given specific rights or permissions of her corresponding system role, one between administrator and innovator.

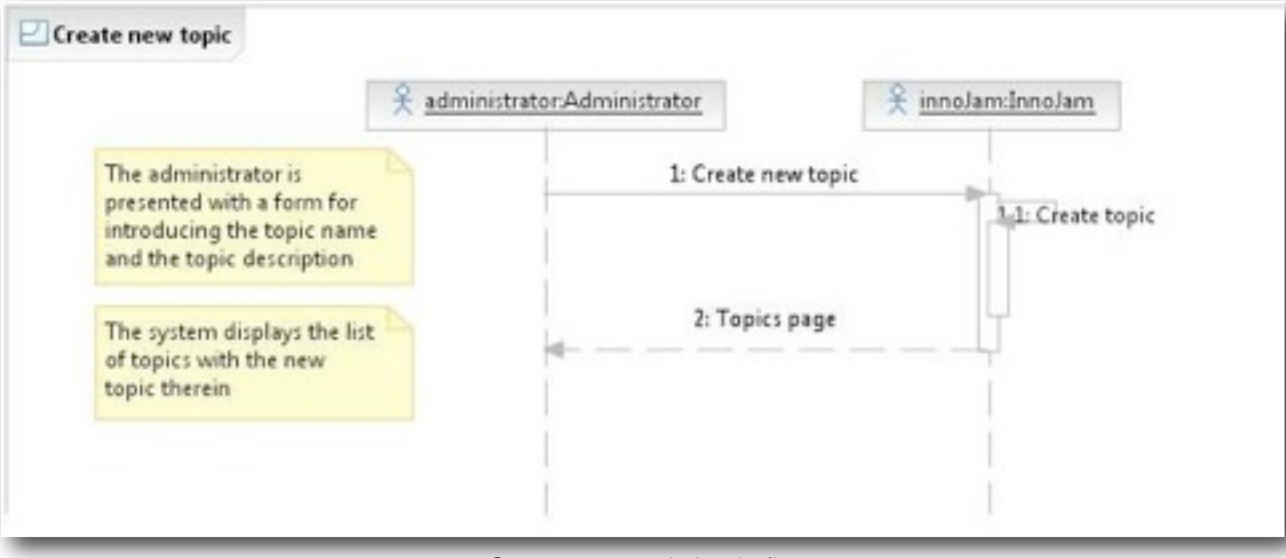


Validate access basic flow

Create a new topic

The administrator role has the ability to create new topics in the system for the innovators to discuss about. In order to do that, she enters proper information such as the topic title and optionally topic description which can contain also background information (this information will be linked or embedded).

After that moment, innovators may start creating discussions and proposals on this just created topic.

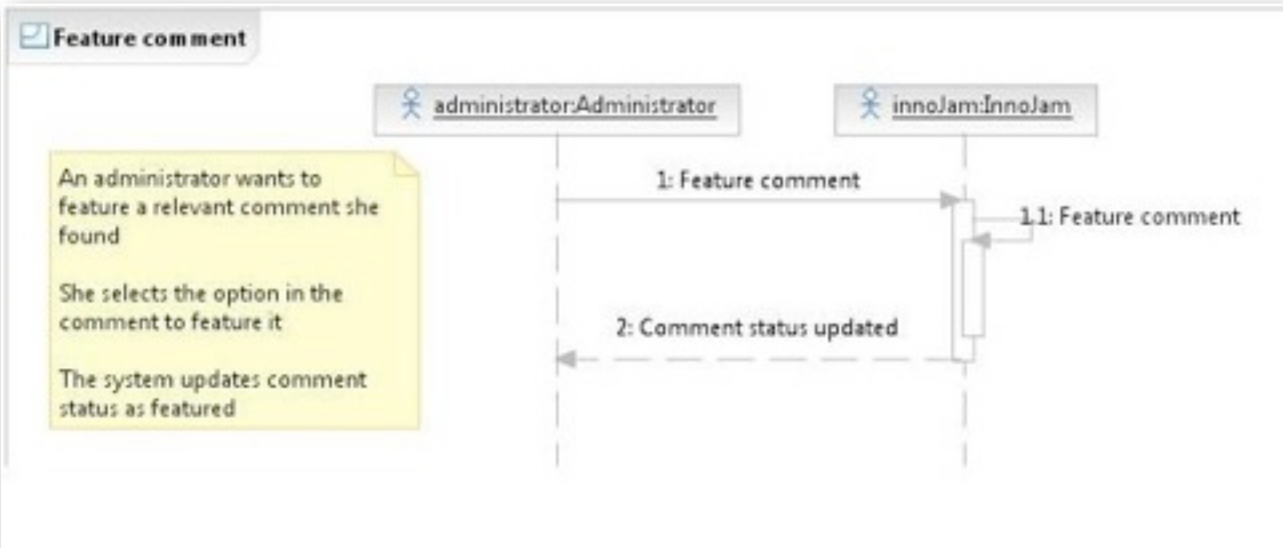


Create new topic basic flow

Feature comments

The administrator or moderator role is also responsible for featuring relevant comments into some prominent space within the InnoJam system.

For performing that action, the user enacting this role will have an option for every comment to be featured. This action, however can be undone at anytime by any user with enough privileges, i.e. one moderator user.



Feature comment basic flow

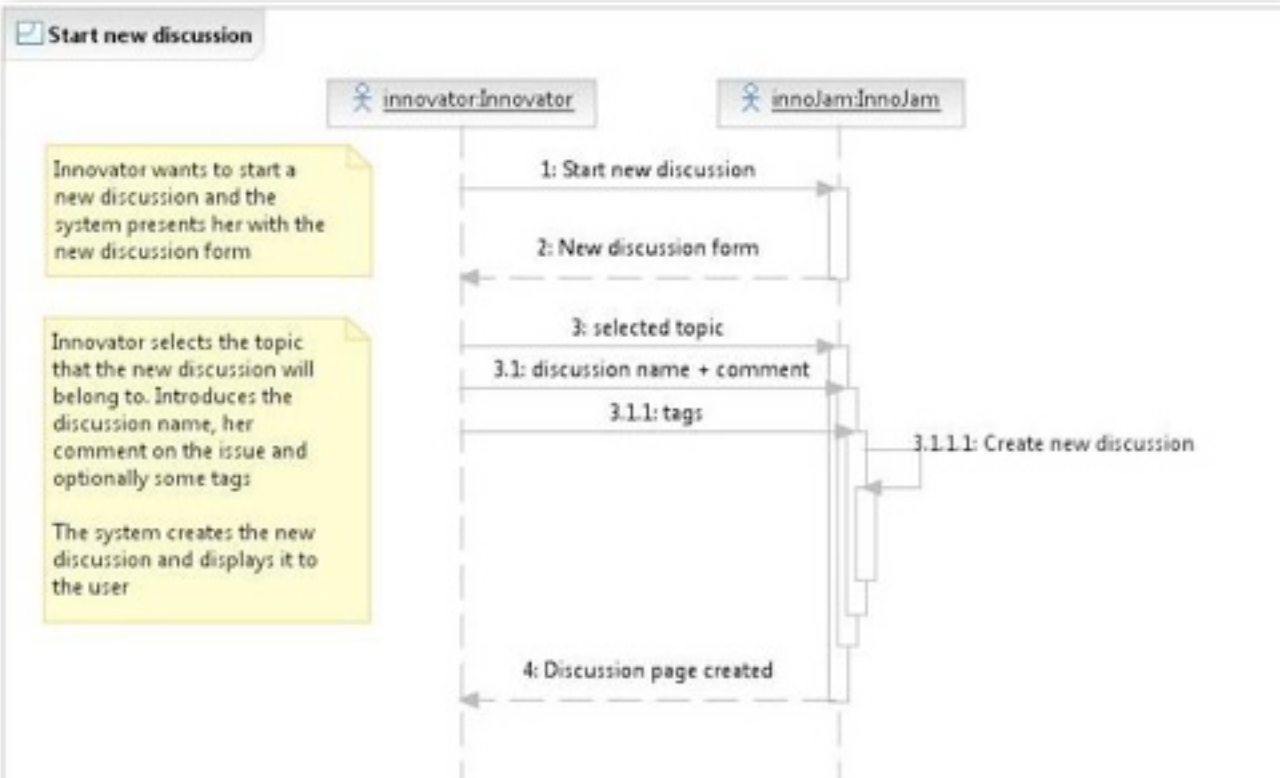
Start a new discussion

Innovators are allowed to create new discussions inside a previously created topic.

By creating a discussion, they may start the development of an idea or getting people together to comment on a certain issue trying to resolve it, etcetera.

Innovator should enter some information about the discussion she is about to create, e.g. it is mandatory to introduce a title or a name for the discussion, which should be descriptive about what is intended to be discussing about, a brief comment on this idea, optionally containing extra information about the topic or the discussion idea, e.g. videos, links, etcetera and optionally at last, introduce some tags that define or serve as keywords for the discussion.

Since that moment, new comments and opinions can be stated by the innovators community.



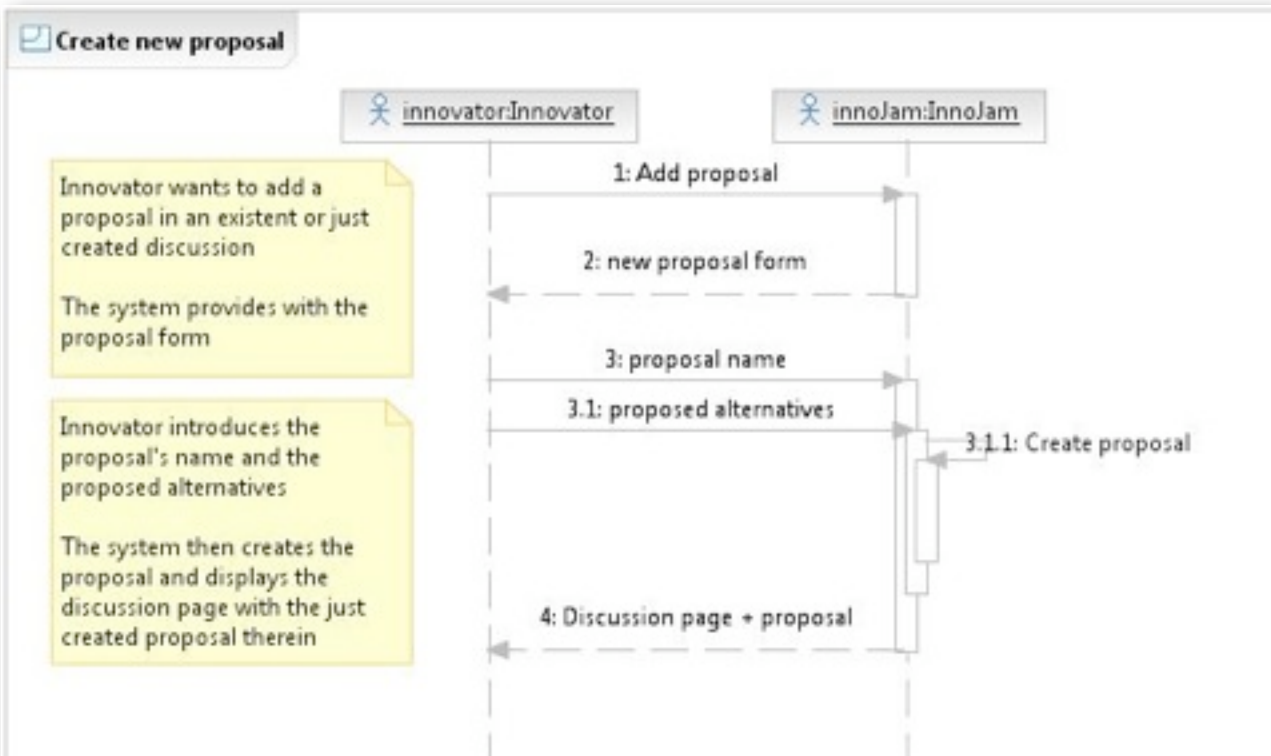
Start new discussion basic flow

Create a new proposal

Innovators are also allowed to create proposals inside a previously created topic, that is, ways of gathering community opinions or preferences about a certain issue.

In this sense, creating a new proposal will consist on some similar steps like creating a discussion: the innovator should enter the proposal name, that is, the idea behind the proposal, introduce some comments or extra information about this proposal and then define a set of alternatives for that proposal so the innovators community will be able to vote on. Additionally, the innovator may also introduce some key concepts that frame the proposal by means of tags.

Just as a newly created discussion, the created proposal is now able to receive comments and opinions from other innovators, and the different alternatives can be voted on.

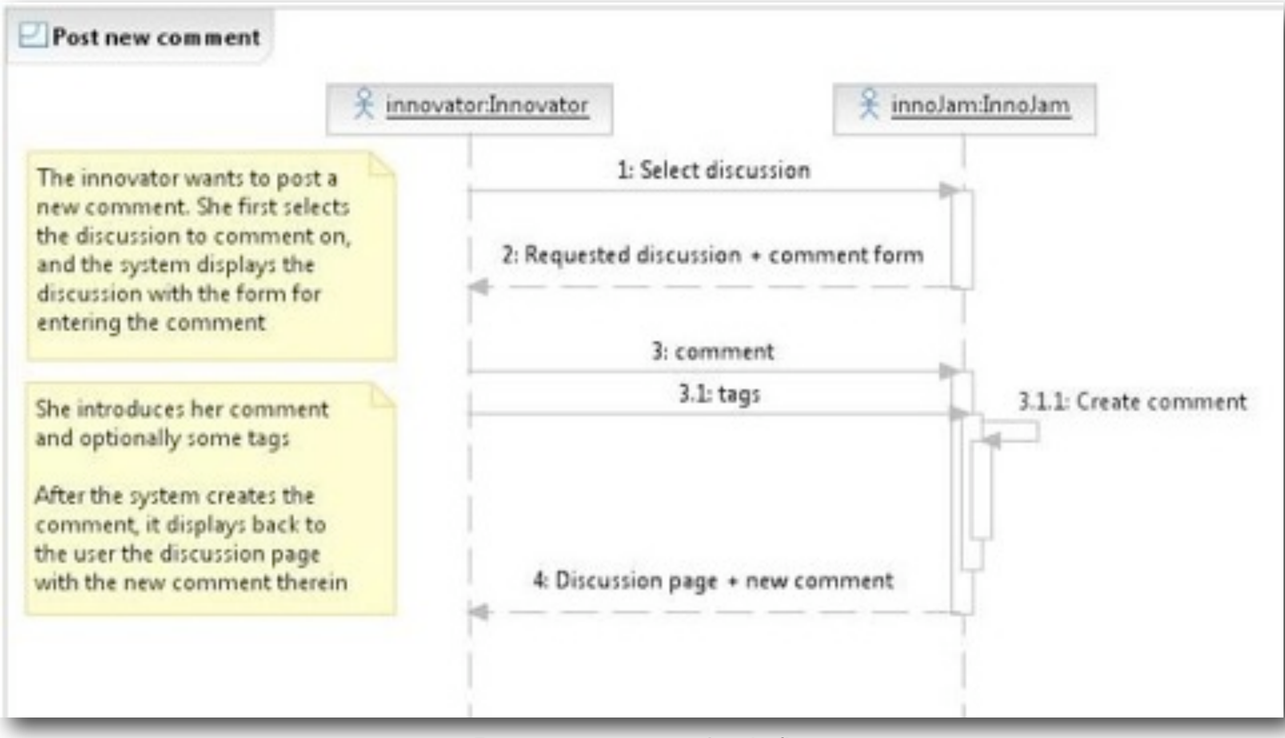


Create new proposal basic flow

Post a new comment

In order to express their opinion on different topics and issues, users can publicly state it inside the flow of a discussion or a proposal of the appropriate subject by posting a comment.

To do so, the participant will select the discussion or proposal to comment on and then will enter her contribution by means of textual or multimedia information. Optionally, the participant may introduce also some tags representing the keywords of her opinion.



Post new comment basic flow

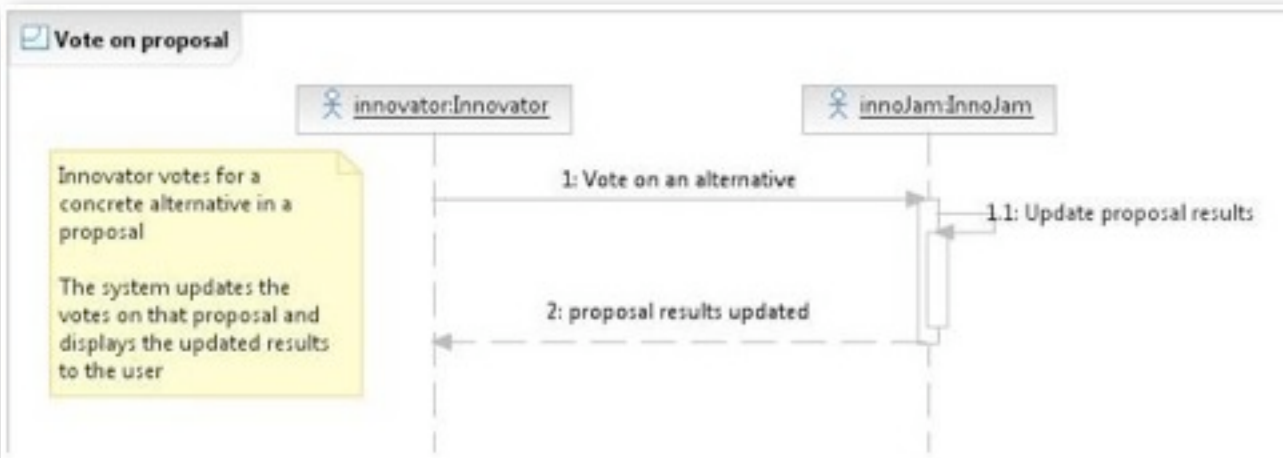
Rate a discussion

Innovators can express their support or discordance to a discussion or an idea by rating the entire discussion with a positive rating or a negative one, respectively.

Rate a discussion basic flow

Vote on a proposal

Innovators can also vote for or against the several alternatives stated within the proposals. The system should give information on the overall results of the proposal.



Vote on proposal basic flow

Report comments

Users can report inappropriate content to the moderators.

Report comment basic flow

Search

Users can run searches in order to find information on their interests. Users are presented with a search form, where they can enter keywords they are looking for.

Some advanced search options can be made available to the user, e.g. filter by the author of the comments or sort the search results by different criteria.

Search basic flow

Get general overview about the event

For getting an overview of the event, users are presented with a special system page where information about the activity of the event can be found.

This information will contain last active discussions, top rated discussions, which topics are being discussed, a map mash-up containing a geo-located version of this activity monitor, a list of comments that have been featured by the moderators and the tag-cloud being produced.

Get overview basic flow

Obtain recommendations

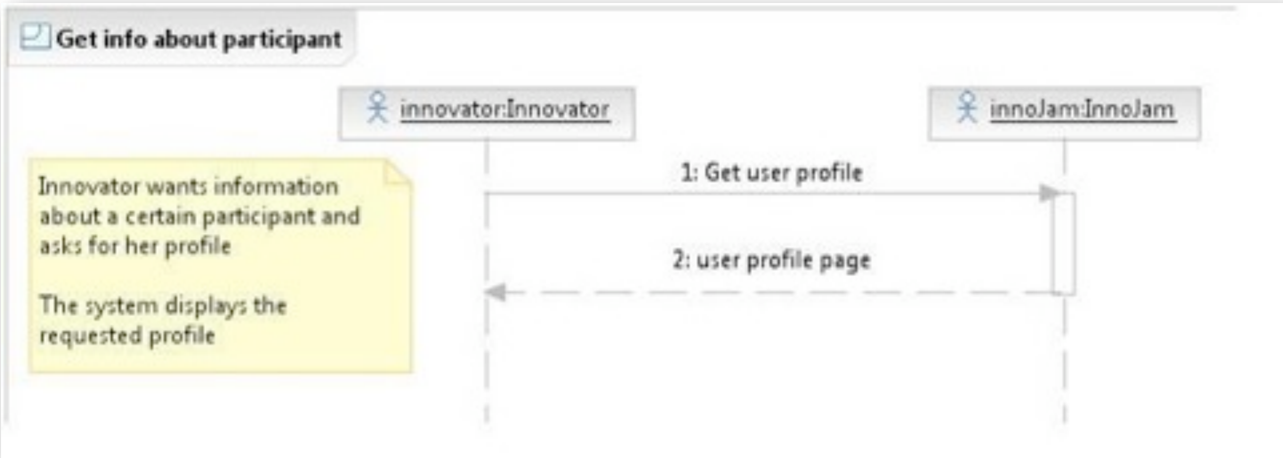
Innovators can get recommended content based on their own profile that the system builds about them.

Obtain recommendations basic flow

Get information about participants

Innovators can get information about other participants by browsing the system to those participants profile page.

This page can contain information about the user such as personal information, their location in a map, their latest contributions, their level of participation within the different topics, etcetera.



Get info about participant basic flow

:)