# SADINA: Adaptive System for Academic News Dissemination

Andrea Barraza-Urbina
Pontificia Universidad Javeriana
Calle 40 #5-50. Bogotá, Colombia
Edificio José Gabriel Maldonado, S.J.
(571) 3208320 ext. 5392
abarraza@javeriana.edu.co

Ricardo Rugeles
Pontificia Universidad Javeriana
Calle 40 #5-50. Bogotá, Colombia
Edificio José Gabriel Maldonado, S.J.
(571) 3208320 ext. 3305
ricardo.rugeles@javeriana.edu.co

Angela Carrillo-Ramos
Pontificia Universidad Javeriana
Calle 40 #5-50. Bogotá, Colombia
Edificio José Gabriel Maldonado, S.J.
(571) 3208320 ext. 5392
angela.carrillo@javeriana.edu.co

## ABSTRACT
Thanks to new information technologies, users can access a large amount and variety of news stories, anytime and anywhere. Nonetheless, current mechanisms for news dissemination do not properly assist users in spotting news articles that could be potentially interesting for the particular user. Commercial applications are developing solutions to address this problem; however, in academic environments, outdated and inefficient practices are still used. Given these reasons, this paper proposes ***SADINA*** (acronym in Spanish of ***S***istema ***A***daptativo para la ***D***ivulgación ***I***nstitucional de ***N***oticias ***A***cadémicas, Adaptive System for Academic News Dissemination, in English). *SADINA* considers information on new stories (*i.e.,* keywords, publication date), users (*i.e.,* type of user) and context (*i.e.,* access device) in order to provide each user with an ordered set of news items according to the user's individual interests.

## Categories and Subject Descriptors
H.3. [**Information Storage and Retrieval**]: H.3.3 Information search and retrieval – *Information filtering, Retrieval models, Search process, and Selection process.*

## General Terms
Algorithms, Design, Theory

## Keywords
News, Information Adaptation, Recommendation Systems, Information Retrieval, Context

## 1. INTRODUCTION
The web environment has proven to be ideal for the mass publication and distribution of content. As a result, traditional systems for news dissemination – channel through which news is delivered – have transformed to take advantage of state-of-the-art technologies, and therefore, users now have instant access to virtually all kinds of news about all kinds of topics. With the creation of web *newspapers*, different information sources have emerged (different in quality, credibility, dissemination channels – *RSS* channels, mailing, webpages –) that produce news stories for many types of audiences. However, with the effects of globalization and the explosion and democratization of information, the limitations of technology in order to respond efficiently to the needs of users in a mass medium for news delivery are increasingly evident.

A user in a web environment typically faces a large quantity and variety of news content, which she/he must analyze and filter to identify interesting stories; all tasks that typically lead to information overload. A possible solution lies in offering services that deliver information to users tailored to their needs, preferences and that consider the user´s context. These services are provided through Information Adaptation [1][2].

Recent trends indicate that the next generation of services for news dissemination should provide tailored content and recommendations. Consequently, several companies have accepted the challenge of providing adapted news services: *Reddit*[25], *Feeds 2.0*[10], *Google News*[12], *Yahoo News*[30], *Flipboard*[11], *Trove*[28] and *LinkedIn Today*[19] (see *section* 5). Moreover, the biggest news networks are initiating projects to offer personalized services, including: *The New York Times* [22], *CNN*[7], *The Washington Post*[29], among others. They have noticed that news recommendation services are not only useful to assist users to find interesting news stories; but also as a *competitive advantage* in the business world by providing added value over rival similar services.

Particularly, in educational settings, traditional channels for news distribution are not enough to capture user loyalty. First, traditional channels do not consider personalized user information or a user's distinct information needs; rather the news offering is oriented to the needs of the academic institution. Second, some institutions maintain antiquated practices which have proven to be ineffective, such as sending generalized mailing to provide news. Third, search services of various academic news portals do not support users in finding the news they need. Finally, it is important that academic institutions act to the latest mobile trend: users now want to read their news in their mobile devices.

In order to respond to the insufficiencies of news dissemination systems, especially in academic environments, this paper proposes ***SADINA*** (acronym in Spanish of ***S***istema ***A***daptativo para la ***D***ivulgación ***I***nstitucional de ***N***oticias ***A***cadémicas, Adaptive System for Academic News Dissemination, in English). *SADINA* considers information on news stories (*i.e.,* keywords, publication date), users (*i.e.,* type of user) and context (*i.e.,* access device) in order to provide each user with an ordered set of news items according to the user's individual interests. As a result, through *SADINA*, users can find relevant news in less time, increasing the level of user satisfaction with the system.

This paper is structured as follows: *section 2* discusses the potential for adaptive systems in news dissemination. In addition, it is also


*iiWAS2013, 2-4* December, 2013, Vienna, Austria.

important to highlight the challenges of incorporating adaptive services, which are identified in *section 3*. Next, in *section 4*, *SADINA* is described: logical architecture, data model, adaptation strategy and the functional prototype. Subsequently, we analyze related works to highlight the advantages of *SADINA*. Finally, conclusions and future works are presented.

## 2. POTENTIAL FOR INFORMATION ADAPTATION IN A NEWS SYSTEM

Given the mass publication of content on the web, it is a major challenge to provide each user with news that best corresponds to their interest and preferences [18]. This challenge can be faced by providing services focused on *Information Adaptation*. As a main advantage, offering each user only relevant information reduces the time spent by the user to search, join, analyze and select content in order to identify the information that is actually interesting to the user. Applications similar to *SADINA* (see *section 5*), focused on the adaptive dissemination of news stories, have found an increase in user satisfaction and loyalty compared to traditional news systems by incorporating adaptive services.

Furthermore, recent trends show that users want to read news in their mobile devices, which allow access anytime and anywhere to the information they need (bearing in mind possible connectivity limitations). The development of mobile applications imposes its own challenges, distinct and more complex than the development of traditional applications given the unique and restrictive characteristics of mobile access devices: screen size, storage capacity, processing capacity, connectivity, among others [5].

In order to mitigate some of the limitations of mobile applications, adaptive services that react and modify presentation or content according to contextual characteristics (*e.g.*, mobile device features, connection type) can be used. Particularly, an adaptive service can be tailored to [5]: (a) *screen size:* it is important to use visual components that enable users to access more information in less space without affecting the usability of the system (b) *type of connection:* the format of news content can be changed according to the user's connection type in order to reduce the amount of transmitted data (*e.g.*, if the user has low connectivity, the system will not present videos, only images).

In addition, adaptive systems are leveraged by contextual information, which can be obtained by means of sensors in mobile devices (*e.g.*, Global Positioning System). A mobile device can capture information associated with user activities, transactions with the system (*e.g.*, clicks), time (time of user connection), location among others [31]. With this information, an adaptive system can characterize the preferences and interests of its users and also deliver location sensitive information (*e.g.*, a user maybe interested in different news stories when at work than at home).

This section argues that the incorporation of adaptive services in news dissemination systems can increase their effectiveness and efficiency. Given these reasons, *SADINA* has been designed as an adaptive mobile application in order to increase user satisfaction. However, it is also important to highlight the challenges that must be overcome in order to incorporate such adaptive services. These will be identified in the following section.

## 3. CHALLENGES FOR INFORMATION ADAPTATION IN A NEWS SYSTEM

This section seeks to highlight some of the challenges and considerations of incorporating adaptive services in a news dissemination system. First, the development and maintenance of an adaptive system is complex, and a cost-benefit study should be carried out before deciding to build one. Benefits must outweigh costs for Information Adaptation to make sense.

Second, when creating adaptive systems it is inevitable to collect user information to provide services, potentially leading to information privacy challenges. *Lee and Cranage* [17] suggest that despite the benefits of adaptive systems, they generate information privacy concerns among users that negatively affect their opinion of the system. However, the value perceived by users from adaptive systems increases when concerns about their privacy are adequately addressed. For example, users are more willing to share information if they feel they have control over it.

Finally, one of the most important challenges faced by adaptive systems for news dissemination – works similar to *SADINA* – is the characterization of both users and news items.

On the one hand, the unique characteristics of a news item must be considered. A news item can be categorized in different topics and is generally structured with a set of keywords. Yet, the product as such is unstructured text. Therefore, if a proper set of keywords is not defined (keywords are usually indicated by an editor), techniques must be used to extract features that describe the news item (*e.g., TF-IDF* [4]). It is important to highlight that on the web most news items are not properly structured.

What's more, the highly dynamic behavior of the news domain must be considered. A news story has a short period of relevance: most stories lose their validity after 24 hours [18]. Because news stories are being constantly created, published and removed; a system for news dissemination should be adaptive and scalable in order to manage a large fluctuating volume of news products.

On the other hand, given the dynamism of the news domain, users have interests that evolve over time [18]. A news dissemination system should be able to react to these changes. Nevertheless, it is difficult to capture the changing interests of users.

This section has highlighted some of the challenges to be faced when creating adaptive news systems. Given the advantages and considering the challenges, the following section proposes *SADINA*: Adaptive System for Academic News Dissemination.

## 4. SADINA

***SADINA*** — acronym in Spanish of ***S***istema ***A***daptativo para la ***D***ivulgación ***I***nstitucional de ***N***oticias ***A***cadémicas, Adaptive System for Academic News Dissemination, in English — proposes a solution to problems associated with information overload in news dissemination systems by incorporating adaptive services. This system will take into account news story information (*i.e.*, title, keywords, header, associated media, description, *etc.*), user information (*i.e.*, age, education, culture, religion, likes, interests, preferences, needs, *etc.*) and contextual information (*i.e.*, access device, location, time, *etc.*) in order to offer each user an ordered list of news items according to the user's interests. This list will be organized considering the following criteria: *(i)* according to the associated *relevance* of the news item given the user: a story will have a high priority if it's relevance, calculated from the user's interests, is high *(ii)* according to the *popularity* of the news item: a news story will have high priority if it is popular within the community.

In *SADINA*, *relevance* is defined as the level of importance or significance of an item (news story) for a user according to her/his interests in a given moment in time. Also, the *popularity* of a news story is defined by the number of clicks and likes the story receives from the community.

This section describes in detail the various components of *SADINA*. In the first place, the characteristics differentiating a dissemination news system for an academic setting with respect to a traditional system are highlighted. In second place, the logical architecture of the system is described. After this, the data model and its role in the adaptation strategy is presented. Finally, we present the functional prototype and the tests conducted to validate the design of *SADINA*.

## 4.1 Academic Environments

Academic environments are characterized by being closed, and abide to a set of specific rules depending on the way the institution manages information. In general, academic environments have different information sources that are centralized according to a set of policies in relation to the official institutional publications. In other words, approved information is consolidated and revised so it complies with a specific publication format and a set of unified guidelines. In this manner, an advantage of working with academic news stories is that the quality, structure and credibility of the story is guaranteed by a known and trusted editor team who filter, standardize, categorize and organize news.

As a disadvantage, academic news bulletins tend to solely consider the intuition of directives of what should be published, and therefore do not really take into account the information needs of potential readers. Furthermore, news presentation is not adapted according to each user but according to what editors believe is a proper news organization. This situation greatly reduces user satisfaction and loyalty to academic news systems because users cannot easily find the news stories they are actually interested in. To make matters worse, news is distributed over outdated channels, such as mailing. Different channels, such as mobile devices, would better respond to user news access trends.

As a case study, this work has been developed for the Pontificia Universidad Javeriana, Bogotá, Colombia. In this institution, information is centralized and governed by the *Office of Communications* who monitors the quality, credibility and editorial guidance of published content. Even though the *Office of Communications* offers a foundation for a unified closed environment, it also leads to publications biased by editorial policies and practices. In addition, the institution uses mailing and a web portal news bulletin to distribute news. Mailing in this institution has proven to be highly ineffective: of 45,000 sent mails a day, 2% of mailing is effective. Similarly, the web portal has approximately 82% of user abandonment. *SADINA* has been proposed as an alternative to news dissemination in academic environments such as the Pontificia Universidad Javeriana. The main objective is to increase user satisfaction when retrieving academic news, while still taking into account the academic institution's goals, practices and restrictions.

## 4.2 Logical Architecture

The logical architecture in *SADINA* is divided in three levels: data level, adaptation level and presentation level. The following sections explain each of the levels in detail.

### *4.2.1 Data Level*

The *Data Level* is in charge of retrieving data from external sources and managing all data for *SADINA.* This level is composed of three components: (1) *News Extractor:* component in charge of retrieving news from the traditional academic news bulleting. This task is carried out autonomously and periodically in order to maintain *SADINA* consistent with the information published on the web by the information source. (2) *User Information Extractor:* component in charge of extracting information about users from the academic institution's user directory (*e.g., LDAP – **L**ightweight **D**irectory **A**ccess **P**rotocol),* (3) *Data Repository:* component that stores the data model (view *section 4.3*).

### *4.2.2 Adaptation Level*

The *Adaptation Level* contains the *Adaptation* component, which executes the adaptation strategy in order to generate for each user a set of news that: is organized according to news relevance (which directly depends on the particular user's interests) and highlights news stories which are currently popular within the community. The *Adaptation* component is also in charge of analyzing user feedback (explicit and implicit) to detect user interests and proceed to build the adaptation model (view *section 4.3.2*) for *SADINA.* This model is saved in the *Data Repository.*

### *4.2.3 Presentation Level*

The *Presentation Layer* contains the *Presentation* component, which interacts with the user through an access device (*e.g.*, mobile client), offering each user adapted services and capturing events of user interactions with the system (*e.g.* user clicks). Even though this layer is specially designed to interact with mobile devices, *SADINA* can also be accessed given a standard Internet Browser. This layer presents to the user the complete set of news stories, which are organized by the *Adaptation* component according to the relevance for each story calculated from the user´s interests. Relevant news for the community (*i.e.*, popular news), are marked with an icon.

In general, the previously described logical architecture can be seen in *Figure 1*. The data model used in *SADINA* is created based on the logical architecture, and is explained in the next section.

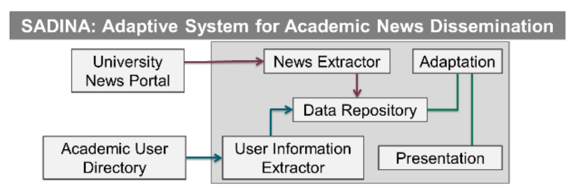


**Figure 1. SADINA Logical Architecture**

## 4.3 Data Model

*SADINA* defines a data model composed of two sub-models: the Domain Model and the Adaptation Model. The domain model is in charge of maintaining the *News Profile*. The adaptation model is in charge of maintaining the *User Reader Profile* and the *Contextual Profile. Figure 2* shows the relationship between the data model and the *SADINA* Logical Architecture. This section will explain both the domain and the adaptation models in detail.

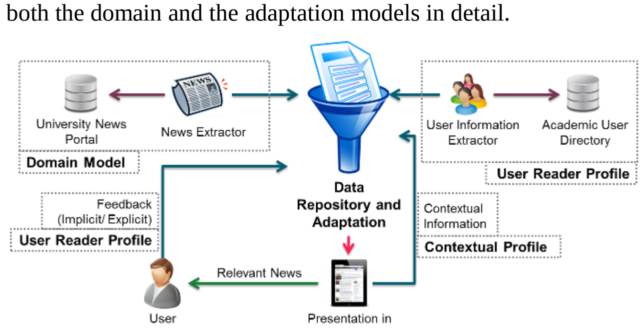


**Figure 2. Relationship between the data model and the SADINA Logical Architecture**

### 4.3.1 *Domain Model*

The Domain Model is in charge of the *News Profile*. The *News Profile* characterizes a news story with information such as: title, header, content of the story, publication date, associated images or multimedia, and description, among others. This information is obtained from the traditional academic news system with the *News Extractor* component. This profile specially highlights the use of keywords to describe the content of each story. A dictionary of pre-defined keywords is defined by the news editor. In the carried out case study (see *section 4.1*), 54 keywords were defined; nonetheless, the model is flexible enough to allow the number of keywords to be reduced or extended.

#### 4.3.1.1 *Acquisition of News Stories*

News stories are acquired by *SADINA* with the *News Extractor* component that runs periodically. This component retrieves news story information from html-based websites that are structured according to known publication patterns. These websites are found in the traditional academic news dissemination system. The information to be retrieved is defined on the Domain Model. Extracted information is placed in *JSON* [14] format and is stored to the *SADINA Data Repository* component.

Before adding news content to the *Data Repository*, the component checks the validity (is the story outdated?) of each story. The validity of news is defined by the editor of the traditional news dissemination system. In this manner, if a story is not currently published, which means it is not within the set of recently extracted stories, it is considered invalid or outdated.

#### 4.3.1.2 *Representation of News Stories*

Within the *News Profile*, the relationship between a news story and its associated keywords is represented with the matrix structure shown in *Figure 3 – Matrix of Current News.*

| | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | … | $w_k$ |
|---|---|---|---|---|---|---|---|---|
| Keywords | 1 | 0 | 1 | 1 | 0 | 0 | … | 1 |

| | | | | |
|---|---|---|---|---|
| idNews 1 | $w_{1,1}$ | $w_{1,2}$ | … | $w_{1,k}$ |
| idNews 2 | $w_{2,1}$ | $w_{2,2}$ | … | $w_{2,k}$ |
| idNews *3* | $w_{3,1}$ | $w_{3,2}$ | … | $w_{3,k}$ |
| … | … | … | … | … |
| idNews *n* | $w_{n,1}$ | $w_{n,2}$ | … | $w_{n,k}$ |

**Figure 3.** ***Matrix of Current News:*** **News x Keywords Matrix**

In the matrix, the first column identifies each news story with a unique identifier. The following columns represent an array that indicates if the news story is associated or not to a specific keyword (1 – the keyword is associated, 0 – the contrary).

### 4.3.2 *Adaptation Model*

*SADINA's* adaptation Model is composed of the *User Reader Profile* and the *Contextual Profile*.

The *User Reader Profile* characterizes a particular user. It seeks to have basic user information such as date of birth, faculty to which the user belongs to, type of user (student, teacher, employee...), among others. On one hand, this information can be obtained explicitly from users or from the academic user directory (*e.g., LDAP*). On the other hand, in order to identify the different information needs of users, their tastes and preferences (one must consider that this information is dynamic); implicit information is captured from the user's transactions with the system (historical data) and the user´s context (*i.e.*, access device).

The *Contextual Profile* characterizes information associated to the access device and the academic institution, which should be considered when offering the adaptive services.

In general, with the adaptation model, services can be enriched to consider user interests and tailor information according to the user's context. The following sections will explain the *User Reader Profile* and *Contextual Profile* in detail.

#### 4.3.2.1 *Acquisition and Representation of the Initial User Reader Profile*

When a user initially starts using *SADINA*, basic user information is retrieved using the *User Information Extractor* component. In order to define basic user information, the system, given a user's academic account credentials, connects with the institution's user directory (*i.e., LDAP*) and extracts data associated with the user, such as: the user's name, date of birth, education level, faculty to which the user belongs to, type of user (student, teacher, employee...). Data retrieved with the *User Information Extractor* is added to the *Data Repository.*

In addition to basic user information, the *User Reader Profile* contains information on user interests. A user interest score is calculated for each of the keywords defined to describe news stories. In this manner, a numerical value is assigned to each keyword that indicates the level of user interest for news stories associated with that keyword. Just as with news stories, the relationship between users and keywords is represented by a matrix (see *Figure 4*). The first column represents a unique identifier for each user. The following columns represent an array that indicates the level of interest of the user for each keyword.

| | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | … | $w_k$ |
|---|---|---|---|---|---|---|---|---|
| Interests | 1 | 5 | 10 | 30 | 40 | 0 | … | 34 |

| | | | | |
|---|---|---|---|---|
| idUser 1 | $w_{1,1}$ | $w_{1,2}$ | … | $w_{1,k}$ |
| idUser2 | $w_{2,1}$ | $w_{2,2}$ | … | $w_{2,k}$ |
| idUser *3* | $w_{3,1}$ | $w_{3,2}$ | … | $w_{3,k}$ |
| … | … | … | … | … |
| idUser *n* | $w_{n,1}$ | $w_{n,2}$ | … | $w_{n,k}$ |

**Figure 4.** ***User Interest Matrix:*** **User x Keywords Matrix**

To retrieve initial information on user interests, users can explicitly modify their *User Reader Profile* within *SADINA* to establish keywords they find the most interesting.

For the case study, a survey was developed (applied to the academic and administrative staff) to find general statistics on user interests according to the user's dependency, type of user, and other features. From findings, demographic profiles were created according to the type of user. These profiles are used as seed information when data on user interests is unknown, specifically with new users.

#### 4.3.2.2 *User Reader Profile Feedback*

User feedback is used to learn user interests from the past interactions of the user with *SADINA*. Feedback is obtained in two ways: (1) *explicit feedback*: the user explicitly indicates her/his interest on a news story with the "*Like"* button or by modifying their *User Reader Profile* to establish which keywords she/he finds the most interesting (2) *implicit feedback:* a user indirectly expresses interest on a story by requesting the full article. When explicit feedback is received for a story, the associated keywords of the news story are granted two points within the user's interest matrix. When implicit feedback is received for a story, the associated keywords of the news story are granted one point within the user's interest matrix. In this manner, *SADINA's* feedback point system was defined to indicate that explicit feedback is twice as important as implicit feedback provided that explicit feedback is much more reliable.

Given that a user has different information needs and context each time she/he accesses a news system, it is natural for the user to also have dynamic interests. Because user interests change over time, a *decay function* is used in *SADINA*. Not only is the *decay function* a response to consider dynamic user tastes, but it is also needed to prevent excessive growth of assigned points in the User Interest Matrix. In general, the following heuristic is applied: the user reduces her/his interest in a topic as the user reads news that is not related to the topic. The core of the *decay function* consists of multiplying points assigned to all keywords by a decay constant (*i.e.*, value lower than one – the case study used 0.95) each time the user decides to read a story.

### *4.3.2.3 Contextual Profile*
Context is defined as the set of elements and their interactions that may affect the execution of a system [9]. Particularly, *SADINA* considers information about the access device and the academic community. On one hand, *SADINA* considers the connection status of the access device and can offer services with or without internet connection based on previously created local news caches. In addition, the screen resolution of the access device is used to tailor presentation to suit the user's requirements. On the other hand, the directives and regulations of the academic institution are also considered in *SADINA* to provide a set of rules that override the adaptation strategy. Particularly in the case study, a rule was needed to indicate that some news stories are of higher priority within the publication. These news stories should be always displayed regardless of the user's interest.

## 4.4 Adaptation Strategy
*SADINA* proposes a strategy that organizes a set of relevant news for each user, based on the integration of the Domain Model and the *User Reader Profile* (view *Figure 5).*

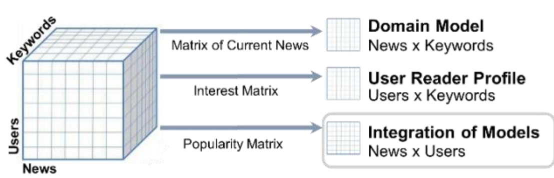


**Figure 5. Integration of Models**

This integration is reflected in the Popularity Matrix, shown in *Figure 7*. In this matrix, each row represents a user and each column a news story. The intersection holds a score, defined as the *Normalized points for the News Story,* which represents how much one user is interested in a particular news story. These scores are calculated using the Matrix of Current News (view *section 4.3.1.2*) and the Interest Matrix (view *section 4.3.2.1*).

*Figure 6* shows an example of how to calculate the *Normalized points for the News Story* for a particular user and a particular news story. In the first place, the *Number of total possible points* a user can assign to a story is calculated. This value is the sum of all points among all keywords in the vector of user interests, found in the Interest Matrix. In the case of the shown example, this value is 11. After this, the *Number of points for the News Story* is calculated. This score is the sum of values found in the user interest vector, but only for the keywords assigned to the news story (indicated in the Matrix of Current News). In the example, this score is 8. This score is the first approximation that describes the user's interest for the news story. Nonetheless, given that all users don't have the same *Number of total possible points*, the *Number of points for the News Story* must be normalized. Scores will be placed in a range of [0-100]. In this manner, as a final step, the *Normalized points for the News Story* is calculated as the *Number of points for the News Story* divided by the *Number of total possible points* multiplied by 100.

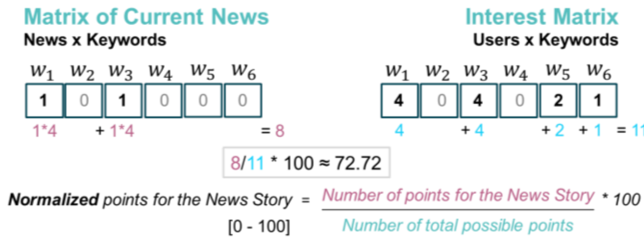


**Figure 6. Calculating normalized points for a news story given a user's interests**

Using the Popularity Matrix, *SADINA* can offer two services that organize news according to: (1) *Service #1*: user relevance (2) *Service #2*: the popularity of stories within the community.

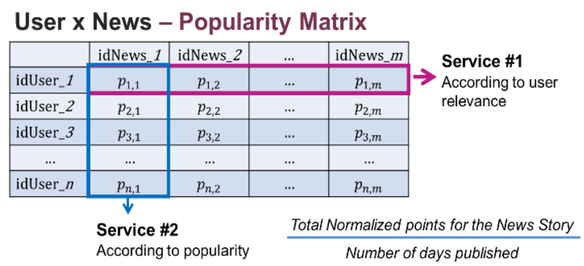


**Figure 7. The Popularity Matrix – Calculating popular news items and relevant news for a particular user**

On one hand, to offer users an organized list of stories, according to user relevance (Service #1), news stories are organized according to the scores calculated given user interests for each story. On the other hand, to offer the service "popular news" (Service #2), each story is assigned a number of points which are the sum of the individual points from all users (how much user's in the community have shown interest in the story) divided by the number of days the story has been published. The number of days the story has been published is considered in order to not bias the popularity of a news story towards older news which, given their time published, could have gained more points than stories recently published. In this manner, *SADINA* offers services in two levels, individual and community-based. *Figure 7* reflects the explained strategy.

## 4.5 Functional Prototype
Given the emerging needs of academic institutions in terms of information and news, and trends in the development and use of mobile devices, *SADINA* was conceived as a mobile solution designed to facilitate access to relevant news stories.

*SADINA* has been implemented under a set of technologies for the extraction, interaction and analysis of user and news information. These technologies are identified in *Figure 8*. In general, technologies used are compatible or based on Java, like the server environment – *Tomcat 7* – and programming languages.

*PhoneGap* [24] was used as a tool for the construction of the mobile application. PhoneGap/Apache Cordoba is an open source mobile development framework that uses standardized *API* allowing developers to create applications in non-device-specific languages such as HTML, CSS, Responsive Design, JavaScript and JQuery.

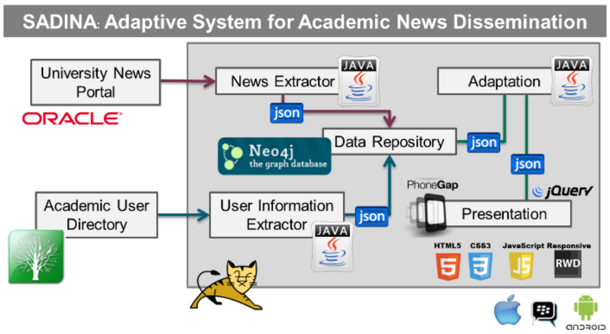


**Figure 8. Employed technology in *SADINA* components**

Data storage is carried out with the graph database *Neo4j* [21], which is later explained in *section 4.5.1*. In general, information on news, users and relationships between data is stored in Neo4j to increase efficiency when developing the adaptation strategy.

For data communication between components, data is formatted to a JSON structure, which facilitates the exchange of information. As mentioned before, basic user information, validation and affiliation is extracted from the institutional user directory *LDAP*. The *News Extractor* component was built by creating a *Web Spider* (specialized software focused on the collection of information based on html publication patterns and other mechanisms [13]) in Java, based on the publishing and editorial patterns of the *Office of Communications* of the Pontificia Universidad Javeriana.

The use of the described technologies allows *SADINA* to be installed as an application in any mobile device compatible with PhoneGap, which means it can be accessed anytime and anywhere. This follows the principles of ubiquitous computing. *SADINA* can also be accessed by any Internet Browser that supports HTML 5.

*SADINA* was developed in three phases: *(1) Analysis stage*: publication patterns and policies of the academic news publisher were studied *(2) Implementation stage without adaptation*: a basic news application, without the adaptation strategy based on the extractor components, was developed. *(3) Implementation stage with adaptation*: the adaptation strategy was added to the basic implementation offering services tailored to the user and her/his context.

**Figure 9**. ***SADINA* prototype in different mobile devices**

The resulting prototype can be viewed in *Figure 9*, where all the published news stories are presented to the user organized by the adaptation strategy; which considers the news most relevant to the user, news most important to the community and the chronological order of the stories.

It is noted that each news story in the main page is presented in the *News Story Box* that contains an associated image and the news story headline. *Figure 10* shows the basic elements of the *News Story Box*. It can be seen that in the upper right corner the *"Like"* button is found. The user can indicate explicitly her/his interest for each story using the associated *"Like"* button. On the upper left corner, a "*Star"* symbol can optionally be found. This symbol identifies stories which belong to the most popular news within the community. In the case study, it was defined that the symbol would identify stories belonging to the group of top ten most popular news stories.

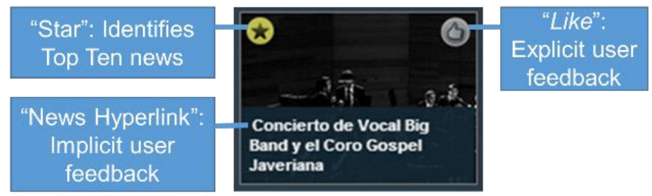


**Figure 10**. **Basic elements of the News Story Box in *SADINA***

Once the application was developed, functional tests were carried out. The following section will explain how Neo4j was used to store data in *SADINA.* Then results for functional tests are shown.

### *4.5.1 Data Storage in Neo4j*

*Neo4j* [21] is a database engine, that supports NoSQL, and allows for storage of a graph data model. Graph databases have several advantages over traditional relational databases, especially when data is very large, interconnected and inconsistent or fluctuating. In the first place, graph databases allow for dynamic data structures while traditional relational databases require a rigid description of the underlying data [26]. Later modifications over the pre-defined structure of the relational database can be costly while graph databases are flexible and easily scalable. In second place, traditional relational databases are ironically not very suited to represent relationships in data [26]. Relationships in relational databases lack semantics: direction, name and associated properties. Relational databases are most suited for tabular structures where relationships tend to exists only as a means to join tables. As information increases in size, expensive table joins critically affect query performance in relational databases. Lastly, real-world data is intuitively represented in graphs rather than tables [26].

In graph models data is represented in nodes and relationships. The graph structure used in *SADINA* is shown in *Figure 11*.

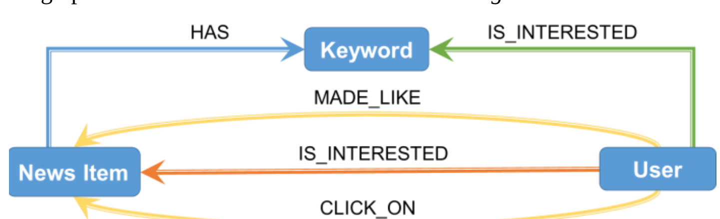


**Figure 11**. **SADINA Data Graph representation in Neo4j**

Here, there is a clear relationship between news and keywords: A news item *HAS* associated keywords. Similarly, there is a relationship between users and keywords: A user *IS_INTERESTED* in a keyword. In Neo4j, relationships can have associated properties, in this case a score indicating how much the user is interested on the keyword. User interactions with news stories are represented by the relationships *MADE_LIKE* and *CLICK_ON*. Lastly, the relationship between users and news items is represented: A user *IS_INTERESTED* on a news item. This relationship also has an associated value that indicates how much the user is interested on the news item.

### *4.5.2 Functional Tests*

Tests carried out for the initial validation of *SADINA* are divided in three types: (1) tests with scenarios (2) presentation tests with different mobile devices (3) tests with users.

Tests with scenarios were taken in two stages: the study of the system with and without adaptation. *Figure 12* shows the system without adaptation, responding to the order given by the editor. Then, tests were carried out with adaptation, observing changes of the system's output according to predefined *User Reader Profiles*. These test profiles were designed with the information retrieved from the survey. *SADINA* showed different and coherent results for the different sample users.


**Figure 12. *SADINA* without adaptation**

After this, functional tests in different mobile devices were carried out to prove that the use of *SADINA* is independent of the access device (see *Figure 9*). The use of techniques associated to "*responsive design*" [20], which enable presentation adaptation, were also tested. In general, *SADINA* was tested in *Android, IOS, BlackBerry* and *Windows.*

Lastly, tests were carried out with users. The testing protocol was based on two surveys:

The first survey was conducted to measure the interests of the different user groups within all the academic community. This survey was answered by 216 individuals: 88 students, 85 employees, 26 teachers and 18 alumni.

The second survey aims to evaluate the level of user satisfaction when using *SADINA* compared to the level of user satisfaction when using the traditional academic news bulletin. This second survey was directed to a group of 16 people, including teachers, students, employees and alumni; who reviewed the results of the application for three scenarios. Each scenario was evaluated on a scale of 1-10 (10 being that results were the most consistent with user expectations and 1 the opposite). These scenarios were defined in accordance to the services provided by *SADINA* (view *section 4.4*)*:* (1) *Service #1:* scenario that offers adaptive services based on information about user interests (2) *Service #2:* scenario that offers adaptive services based on information about the interests of a group of users (3) *Ordered according to the editor:* scenario without adaptation where news is delivered according to an order chosen by the editor.

Test results showed that having more data in user profiles allowed *SADINA* to better enrich the service "news delivery," and thus the user manifested higher satisfaction when using the system. *Figure 13* presents user satisfaction results for the second survey.

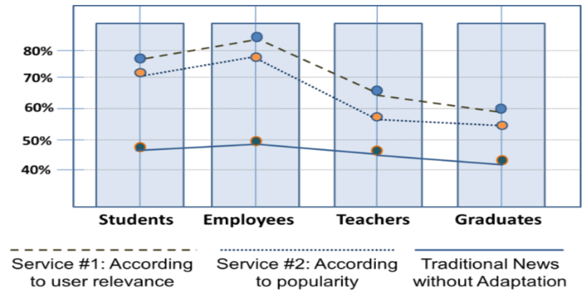


**Figure 13. User Satisfaction Index Survey Results**

In particular, the satisfaction rate associated to news organized by *SADINA* for students and employees (population with greater amount of captured initial information from the first survey) stood above 70%. In contrast, the satisfaction rate for teachers remained on average within the 64%. Results showed that news organized by the editor was considered less valuable for users, obtaining only 47% of acceptance from evaluators.

In synthesis, tests showed that the degree of user satisfaction depends on the amount of information obtained for each user group profile from the first survey. In other words, *SADINA* offers better services as it obtains more user input. As a result, data about teachers and alumni needs to be refined. In general, *SADINA* presented an organized list of news tailored to various user groups, and showed a higher rate of acceptance over the traditional model used for the academic news bulletin. It is expected that the user satisfaction rate for *SADINA* will increase as user transactions and feedback is received.

## 4.6 Challenges when creating *SADINA*

When designing and developing *SADINA*, specifically for the case study, several challenges were overcome. Some of the most relevant are mentioned in this section.

Given security related issues, *SADINA* could not access the academic news repository directly. Because of this, the *News Extractor* component was designed as a Web Spider, in order to retrieve news from a publicly available source. It is important to highlight that the institution's *LDAP* was already publicly available and could be accessed by the *User Information Extractor.*

The applied algorithms had to be very efficient in order to analyze adaptation information and offer results in real time.

The academic institution's culture on the use of keywords had to be re-enforced. The editor was not used to the discipline of adding keywords from a dictionary of words. However, if keywords are not added correctly, *SADINA* will not be able to function properly.

Future and continuous validation of the system is still a challenge, given that *SADINA* will offer better services given time and user transactions with the system. This is a long-term challenge. However, in preliminary testing scenarios, results were positive.

This section has explained *SADINA* in detail. The following section analyzes similar works and compares them to *SADINA* to show the advantages of the proposed system.

## 5. RELATED WORKS

This section seeks to provide an overview of academic works related to adaptive news dissemination. Despite numerous existing research projects, it was not possible to find a work solely focused on *academic environments*. For this reason, works were selected without considering the domain they were intended for. Also, given the commercial success of adaptive systems for news dissemination, some commercial applications are also analyzed. Finally, a synthesis of the considered works is presented.

*NewsWeeder*[16] uses collaborative filtering to discover the interests of each user, based on the interests of all the other users of the system (not just similar users). Through this approach, it is not necessary for a user to explicitly describe her/his interests. Furthermore, *NewsWeeder* applies a filter based on the content of each story to determine which are relevant to the user. Therefore, the result is based on a hybrid approximation, and is obtained from an average of the predictions obtained from both filters [4]. However, in [16] only the content-based filter is described.

*Grouplens*[15] is a collaborative filtering system for Newsgroups. The system allows its users to observe the predicted ratings generated by the system for each news story and choose which news the user wants to read. In addition, the user can evaluate a story with a score within a five rating scale. As any collaborative system, *Grouplens* is based on the heuristic that users with similar tastes in the past are very likely to have similar tastes in the future [4]. From user evaluations, *Grouplens* found that once users invest time in the system, they receive better results and are likely to continue using it [15].

*P-Tango*[6] is a hybrid recommender system, where the estimated rating for each news story is obtained from a weighted average of collaborative predictions and content-based predictions. As more users evaluate a story, the weight of the collaborative predictions associated to the news story increases, avoiding the first-rater problem – in purely collaborative systems, first evaluations tend to skew future evaluations (*e.g.*, if the first user to rate a news item did not like it, then it would be difficult for the item to get recommended) –. Furthermore, predictions are calculated for each user, avoiding the gray sheep problem – refers to users whose views do not always agree or disagree with any group of people and therefore do not benefit from collaborative filtering – [4]. The user profile is divided into sections that correspond to the categories and keywords of news stories. The content-based filter achieves an association between the keywords that describe a story and the keywords that make up a user's profile. A user can set her/his preferences in her/his profile by selecting a set of news categories that represent her/his interests and keywords that are most relevant for each category. This way, content-based predictions calculate the relevance of a story for a user from the distance between the news item's keywords and the keywords found in the user profile, using the *Overlap Coefficient* [8].

*Daily Learner*[3] is a platform that provides adaptive access to news provided by *Yahoo News* [30]. The system models the interests of each user in different categories, based on explicit and implicit feedback. *Daily Learner* is flexible enough to quickly adapt to the changing needs and interests of its members (the needs of users change due to their interactions with the system [4]). To achieve this, a model to represent a user is chosen which consists of two parts: user long-term interests and user short-term interests. To determine if a user is interested in a story, the system applies both models in a sequential manner (first short term model). The short term model only considers the *N* most recent news stories. Learning a short-term model that considers only the most recent news observed by the user allows the model to fit more quickly to dynamic user interests. The long-term model seeks to represent general user preferences.

The *Google News*[8] system manages millions of users, news stories, and information sources. The authors emphasize that current solutions are not appropriate to respond to their unique needs, and thus, they propose novel and scalable algorithms. Recommendations are aimed at users who are logged on to the system, and only implicit positive feedback is considered: a visit to a news story accounts as positive feedback. Thus, the ultimate goal is to present the user with recommendations based on their click history and the community's click history. However, given that new users do not have a click history, the proposed solution is not feasible to recommend news to these users or in general to represent dynamic user preferences. To solve this, an algorithm based on the co-visitation between stories is used. Co-visitation occurs when two stories are visited by a user in a given interval of time. With co-visitation information, a user can be shown stories that were visited along the news that she/he is currently viewing.

*PPNews*[31] presents news to users in a proactive and personalized way, using hybrid P2P technologies. Specifically, the system provides push and just-in-time services to mobile users considering both their contextual information and the content of news stories. The information needs of a user are estimated based on a Bayesian Network. An Analytic Hierarchy Process (AHP) is developed to estimate the possible relevance of a story for a user.

*O'Banion et al.* [23] present an approach to offer personalized news recommendations using the content of news stories and social-media (Twitter) as information sources to estimate user interests. Techniques from Information Retrieval are used to represent tweets and users as a hierarchical collection of news themes. The system models the user using information obtained implicitly from their social interactions on Twitter: the user's Twitter conversations serve as evidence of their preferences and interests. In the first place, a set of themes are previously defined by editors in order to organize news content. Secondly, instances of these themes are searched for in the user's Twitter conversations. In this manner, the user profile identifies categories and tags from Twitter conversations which are assigned a weight that represents the user's interest level associated with the topic. The system also considers that the user's interests evolve over time, and therefore gives more weight to recently detected tags. Additionally, the system uses an exponential decay function based on the last day the tag was mentioned on Twitter. The biggest benefit of this approach is that the process used to model a user does not depend on the contents of a particular news provider.

Currently, there are many commercial applications that adjust their services to provide personalized information to each user based to their interests. The following briefly describes some of the most relevant applications.

*Reddit*[25]*: Reddit* is made up of a collection of entries organized as a bulletin board. The entries are organized by areas of interest or *reddits*. Different users can vote (upvote / downvote) per entry and the page is set according to the tastes of the community.

*Feeds 2.0*[10]: is a system that aggregates *RSS* channels. *Feeds* prioritizes incoming information according to user interests. The system considers for each user implicit (clicks) and explicit (a user can indicate her/his interest on a story) feedback. To recommend to users new entries, the system takes into account the interests of the user community using collaborative filtering.

*Yahoo News*[30]: aggregates news from various sources and categorizes them. The system receives explicit feedback from its users (like/dislike a news story). *Yahoo News* combines the view of the user community and its own editors to organize the presentation of their news. Among the most important services offered by the system are: a user can follow the activity of his friends, section "*News for you*", section of the most popular news, the user can customize his preferences through his "*My Y!*" page.

*Flipboard*[11]: indicates as its mission to allow users to discover and share content in a simple, meaningful and aesthetic way. *Flipboard* collects information from social media and other web sources and displays news with a magazine format, allowing users to flip between the news. A user can build and organize her/his customized magazines, add content, share it, indicate her/his taste and connect to social media.

*Trove*[28]: is a personalized service that aggregates over 10,000 news blogs and pages. *Trove* has a set of editors who are dedicated to creating various news channels organized by news categories. *Trove* collects information on user interests to build the user a home page with the most relevant news channel for the user. Information is collected of user interests from: Facebook, the user's click record, what the user shares, the sources of information that the user values and the themes that the user is interested in. The user can also explicitly add and remove channels from her/his homepage.

*LinkedIn Today*[19]: is a social product that aggregates business related news, identifies the most relevant news according to industries and offers them to the different users based on the information of their professional profiles and the groups to which each user belongs to. To identify whether a story is relevant, *LinkedIn Today* considers various criteria such as: user opinions, user views, user shares, the professional profile of users that interact with the news story, the industry to which the story belongs to, among others. In addition, *LinkedIn Today* works with a staff of editors that can also categorize news as relevant or not.

In conclusion, this section presented a review of works related to adaptive systems for news dissemination. It is important to highlight that not only recent works were considered; works historically relevant were also reviewed. For example, *Grouplens*[15] is one of the most cited research articles as it was one of the first works in collaborative filtering. Similarly, *Daily Learner*[3] is one of the first solutions to consider user dynamic preferences.

In *Table 1* a clear comparison between attributes and functionalities of the presented related works (except for commercial applications) and *SADINA* can be viewed. The symbol '+' indicates the criteria is included, the symbol '-' indicates criteria is not included, the symbol '?' indicates that it is not clear if the criteria is included.

From the analysis of related work it is observed that: (1) given advances in technology, recent solutions began to slowly consider implicit information as valuable feedback for a user profile. (2) Few academic solutions are targeted at mobile devices. However, it was found that various commercial applications are aimed at mobile devices responding to new trends and user needs. (3) Given the social-media explosion, and the valuable information that social-media holds to be added to a user profile, recent research and commercial applications have incorporated social-media as an essential part of their services. (4) It is found that various commercial applications still consider as part of their services the influence of human editors. (5) The majority of solutions consider explicit user feedback. It is determined that this type of information is more reliable than implicit feedback, though it implies an additional effort for the user. (6) Recent solutions observed the need to consider dynamic user preferences to enhance the results of the adaptive system. (7) Even if a hybrid methodology (both content-based and collaborative-based) is desirable, few projects propose it. One possible reason is the complexity of developing a hybrid solution. A hybrid solution is desirable because the disadvantages of a collaborative solution are mitigated with a content-based solution, and inversely. (8) It is highlighted that *SADINA* meets all criteria and is specifically focused on an academic setting.

**Table 1. Comparison of Related Works**

| | **[16]** | **[15]** | **[6]** | **[3]** | **[8]** | **[31]** | **[23]** | **SADINA** |
|---|---|---|---|---|---|---|---|---|
| Publication date | 1995 | 1997 | 1999 | 2000 | 2007 | 2010 | 2012 | 2013 |
| News Domain | | | | | | | | |
| Generic | + | + | + | + | + | + | + | - |
| Specific | - | - | - | - | - | - | - | + |
| Type of Solution | | | | | | | | |
| Web | + | + | + | + | + | + | + | + |
| Mobile Application | - | - | - | + | - | + | - | + |
| Adaptation Characteristics | | | | | | | | |
| Considers news story content | + | - | + | + | - | + | + | + |
| Considers similar users | ? | + | + | - | + | ? | - | + |
| Considers contextual characteristics | - | - | - | - | - | + | - | + |
| Considers user interests | + | + | + | + | + | + | + | + |
| Reacts to changes in user interests | - | - | - | + | + | - | + | + |
| Considers user explicit feedback | + | + | + | + | - | + | - | + |
| Considers user implicit feedback | - | - | - | + | + | + | + | + |

# 6. CONCLUSIONS AND FUTURE WORK

It is truly of great value to have constant availability of updated last minute news stories, thanks to new information technologies. However, availability is useless without the tools that support users to find the news that they are really interested in.

News delivery channels, particularly in academic environments, do not properly assist users in the task of identifying relevant news stories. As a solution, this paper introduces *SADINA*, an adaptive system for the institutional dissemination of academic news. *SADINA* presents an organized set of news to its users to support them in the task of finding, in less time, the news that best fits their interests. To achieve this, *SADINA* considers information on new stories (*i.e.,* keywords, publication date), users (*i.e.,* interests, user type) and context (*i.e.,* access device). With *SADINA*, an increase in user satisfaction and loyalty is achieved.

In this paper, the advantages and challenges to face when creating adaptive systems for news dissemination were explored. In addition, a set of related works were analyzed. Even though these works were not specifically focused on the academic domain, they served as foundation to suggest *SADINA*, an alternative solution for adaptive news dissemination to be applied in an educational setting. *SADINA* is presented in detail and a special focus is given to the development of a functional prototype in order to validate the system.

As future work, we propose to continue to leverage the use of captured information (explicit and implicit) to provide new adaptive services. For example, the service "*others like you also liked ...*" could be provided, in which news stories are offered to users if they have been liked by similar users. In this manner, *SADINA* can be extended with collaborative filtering techniques. In addition, services offered could include co-visitation techniques which associate news stories that have been visited together. For now, *SADINA* includes *pull* services (user-initiated) but there is great potential to incorporate *push* notifications (initiated by the system). Finally, it is important to carry out long term studies on user acceptance to capture stronger results.

Lastly, it is important to highlight that *SADINA* can serve as foundation for similar works. For example, the model proposed by *SADINA* establishes the foundations for *NUNCIUS* (*Nuncius* is a term in Latin which means "*Messenger*")[27], a notifications model for nomad users. *NUNCIUS'* main goal is to focus the notification process (information distribution) on the user, supported by her/his context and allowing feedback from users to improve the offered information in notifications. *NUNCIUS* provides users with notifications from diverse information providers, and as *SADINA* does with news, *NUNCIUS* offers each user notifications relevant to the particular user.

## 7. ACKNOWLEDGEMENTS

This project was developed at the Pontificia Universidad Javeriana, Bogotá, Colombia with the support of the Office of Communications (Dr. Arritokieta Pimentel and Dr. Mauricio Ocampo), the Department of Information Technology (Eng. Luis Francisco Martinez) and the Master Program of the Computer Science Department.

## 8. REFERENCES

[1] Arias-Baez, M. P., Torres-Ribero, L. G., Carrillo-Ramos, A., Pomares, A., and González E. Adaptation for enriching services taking into account mobile contextual features. In: Proc. of the 9th Int. Conference on Advances in Mobile Computing and Multimedia, 236–239, (2011).

[2] Billsus, D. and Pazzani, M. J. Adaptive News Access. In: The Adaptive Web, LNCS 4321, Springer, 550-570 (2007)

[3] Billsus, D. and Pazzani, M. J. User Modeling for Adaptive News Access. In: Journal of User Modeling and User-Adapted Interaction, vol 10, no. 2–3, Kluwer Academic Publisher, 147–180 (2000).

[4] Borges, H., and Lorena, A. C. A Survey on Recommender Systems for News Data. In: Smart Information and Knowledge Management, Springer, 129-151 (2010).

[5] Carrillo–Ramos, A., Kirsch, M., Villanova, M, Gensel, J., and Berbers, J. Collaborating Agents for Adaptation to Mobile Users. In: Collaborative and Social Information Retrieval and Access, IGI Global. 250-276 (2009)

[6] Claypool, M., Anuja G., Tim M., Pavel M., Dmitry N., and Matthew S. Combining Content-Based and Collaborative Filters in an Online Newspaper. In ACM SIGIR Workshop on Recomender Systems, ACM (2009)

[7] CNN Acquires Zite. Retrieved: Feb 2013: http://cnnpressroom.blogs.cnn.com/2011/08/30/cnnzite/.

[8] Das, S., Mayur D., Ashutosh G., and Shyam Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. In: Proc. of the 16th Int. conference on World Wide Web, ACM, 271-280 (2007).

[9] Dey, A., Abowd, G. Towards a Better Understanding of Context and Context-Awareness. In: Proc. of the 1st Int. Symposium on Handheld and Ubiquitous Computing. LNCS, Springer, 304-307 (1999)

[10] Feeds, Retrieved: Feb 2013: http://www.feeds2.com.

[11] Flipboard, Retrieved: Feb 2013: http://flipboard.com/.

[12] Google News, Retrieved: Feb 2013: https://news.google.com/.

[13] Hong, C., Chen, C., Chiu, C. Automatic extraction of new words based on Google News corporation for supporting lexicon-based Chinese word segmentation systems. In: International Journal of Expert Systems with Applications, vol 35, no. 2, 3641-3651 (2009)

[14] JSON, Retrieved: August 2013: http://www.json.org/

[15] Konstan, J. A., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. GroupLens: Applying Collaborative Filtering to Usenet News. In: Communications of the ACM, vol 40, no. 3, 77-87 ( 1997)

[16] Lang, Ken. NewsWeeder: Learning to Filter Netnews In: Proc. of the 12th International Machine Learning Conference, Morgan Kaufmann, 331-339 (1995).

[17] Lee, C., and Cranage, D. Personalization–privacy paradox: The effects of personalization and privacy assurance on customer responses to travel Web sites. In: Tourism Management, vol. 32, no. 5, Elsevier, 987–994. (2011)

[18] Li, L., Wang, D., Zhu, S., and Li, T. Personalized News Recommendation: A Review and an Experimental Investigation. In Journal of Computer Science and Technology, vol 26, no. 5, Springer, 754–766. (2011)

[19] LinkedIn Today. Retrieved: Mar 2013: http://www.linkedin.com/today/.

[20] Marcotte, Ethan: Responsive Web Desing, Retrieved: Jun 2013, http://alistapart.com/article/responsive-web-design.

[21] Neo4J, Graph DataBase, Retrieved: Jun 2013, http://www.neo4j.org.

[22] NY Times to add active personalization of article recommendations. Retrieved: Feb 2013: http://www.poynter.org/latest-news/mediawire/141867/new-york-times-to-add-active-personalization-of-article-recommendations/.

[23] O'Banion, S., Larry B., and Kristian H. Social Media-driven News Personalization. In: Proc. Of the 4th ACM RecSys workshop on Recommender Systems and the social web, ACM, 45-52 (2012)

[24] PhoneGap, Retrieved: Jul 2013: http://phonegap.com/.

[25] Reddit, Retrieved: Feb 2013: http://www.reddit.com.

[26] Robinson, I, Webber, J, Eifrem, E. Graph Databases. O'Reilly Media INC, 2013.

[27] Rugeles, R. *NUNCIUS*: Adaptative Notifications System for Nomadic Environments. Final Project for Master Degree in Computer Science. Pontificia Universidad Javeriana, Bogotá, Colombia (Work in Progress)

[28] Trove, Retrieved: Feb 2013: http://www.trove.com.

[29] Washington Post Tests Personalized News Program. Retrieved: Feb 2013: http://techcrunch.com/2012/02/23/washington-post-personal-post/

[30] Y!News, Retrieved: Feb 2013: http://news.yahoo.com/.

[31] Yeung, K. Fung, y Yanyan Y. A Proactive Personalized Mobile News Recommendation System. In: Proc. of Developments in E-systems Engineering, IEEE, 207-212 (2010).