Creating Usage Sensitive Hierarchical Knowledge Structures
Computer Science Honours Project Report

Ian Saunder
Department of Computer Science
University of Cape Town
Rondebosch, 7701
Cape Town, South Africa

+27 72 638 2997
isaunder@cs.uct.ac.za

Project Supervisor: Dr. Anet Potgieter

ABSTRACT

The Internet has been at the center of an era of innovation that has characterized the last decade, and in many ways is responsible for the success of some of today's most influential technologies. From its inception as a tool for the United States Department of Defense to its present manifestation as a worldwide communications medium, the Internet continues to relentlessly proliferate and expand to all corners of the globe as it endeavors to deliver information to everyone, everywhere.

This extensive growth, however, has single-handedly brought with it several challenges which threaten to transform the Internet into an unstructured mass of unintelligible data, eventually rendering its contents overwhelming to users and meaningless to machines. Search engines have so far done well to make sense of most of this data, and remain the user's primary interface to the Internets vast repository of information. However, the sheer quantity of data to be processed is quickly rendering these keyword-based retrieval mechanisms as inadequate, signaling the need for additional technologies to aide in information classification and retrieval.

Innovative and novel means of data categorization are required to help augment the existing techniques that attempt to produce useful semantic data. One such technique is proposed within this report, and is presented alongside experimentation and tests that provide evidence that provide insight into its merits and failings.

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

While the Internet has proven to be an invaluable mechanism to facilitate human-to-human communications, it presently fails to effectively support the interactions between machines and humans, as well as between machines themselves [1]. In its present manifestation, the Internet is a myriad of inter-connected digital resources which are usually convenient for humans to read, but at the same time almost impossible for machines to interpret. While users are content to view information through informal text and visual imagery, machines find meaning by delving below this abstraction in search of more formalized semantics and structures. Although standards such as eXtensible Markup Language (XML) have provided a means to digitally express user-defined information structures, they nevertheless fail to impart any sense of meaning to the data they express, and leave the user to manually separate the significant from the irrelevant. The aforementioned failures of the Internet at present suggest the need for advancements that will render information on its networks as meaningful to machines as it is to humans. This extension of the Internet, termed the Semantic Web, enables information to be given well-defined meaning, better enabling computers and people to work in co-operation [2]. It is with these challenges in mind that the project in question be presented.

Ontologies and similar information structures have been developed by experts in an attempt to fashion meaning out of the vast expanse of information that litters the digital realm. However, while these mechanisms of conceptualization undeniably create meaning where little existed before, they nevertheless do so at costs which are proving to be too great. More recently, however, casual Internet users have provided a means of collaborative classification by tagging, generating flat structures known as folksonomies. While this novel approach certainly does not provide a complete solution, it nevertheless strengthens the connection between the Internet and the machines that it depends upon.

The primary goal of the project around which this report is based is to investigate novel mechanisms that are able to create data that is present and current. While the managed hierarchical data models that are mentioned above serve several notable purposes, they are inherently static in nature, and are limited in their ability to adapt to changing perceptions. This shortcoming ensures that numerous relationships that they communicate are often dated.

In an attempt to investigate techniques that are able to partially remedy this hindrance, manual and automatic means of textual categorization are combined to produce a system whose output will express a current interpretation of the relationships between the terms of a limited vocabulary. Positive results will be epitomized by semantic data that is both current and useful to individuals and organizations that rely on data categorization and retrieval.

2. BACKGROUND

Metadata is often popularly defined as being *data about data* [3], and serves to provide information about documents, images, and other resources. It exists primarily to facilitate the location and access of data by identifying and attempting to organize items based on their intellectual content [3].

Metadata has long co-existed exclusively alongside professionals who possess the expertise to impart these labels with data according to pre-defined rules and schemes. While these practices generally ensure that the created metadata is of the highest quality, they are costly in terms of both time and effort, and are unable to scale and keep pace with the relentless proliferation of modern-day information structures, such as the Internet. Consequently, innovative alternatives to professionally-created metadata have been popularized by the proliferation of social bookmarking practices which encourage users to assign tags to resources, thereby independently creating metadata of their own. While these novel approaches to metadata creation have shown great promise, they nevertheless have several shortcomings associated with them.

These different approaches to data annotation are discussed within this paper, with an interesting compromised posed as forethought towards the end of the paper.

2.1 Managed Hierarchical Structures

2.1.1 Taxonomies

A Taxonomy is a classification structure that is hierarchical in nature in that the information that it represents is divided into progressively narrower and more specific categories [4]. As such, types are often defined explicitly according to parent-child relationships by professionals who possess expertise within a particular domain. While this approach to categorization has proven to produce precise and accurate knowledge structures, the rate at which information repositories such as the Internet change ensures that the development and maintenance of taxonomies is a costly exercise. The pre-defined categories that constitute such structures are often insufficient to accurately represent the vocabulary of dynamic environments, resulting in exceedingly large items being classified under *other*.

2.1.2 Ontologies

An ontology is defined as a formal specification of a shared conceptualization [5]. They extend the idea of a taxonomy in the sense that they explicitly define the relationships between the different entities within a domain beyond those expressed by simple hierarchical arrangements, thereby transforming the structure into a directed graph [6]. While ontologies have been regarded as central to the idea of the Semantic Web, their complexity and exclusivity ensure that extremely high costs are associated with their development and maintenance. These difficulties have served as the motivation behind the adoption of consequent popularity of collaborative tagging.

2.2 Unmanaged Flat Structures

Flat structures are inherently one-dimensional, and are devoid of the hierarchical arrangements and associations that characterize both taxonomies and ontologies [7]. As such, there exists little information that imparts any sense of categorical superiority between elements; a problem that is magnified by the lack of a central authorative figure to oversee the development and maintenance of the structure. Although these obstacles may appear to render these mechanisms of knowledge representation as ineffective, the recent trend of social bookmarking has proven these attempts of annotation to be not only plausible, but also useful.

2.2.1 Tagging

A tag, as it relates to information technology, is generally a keyword that is associated with or assigned to a resource [8]. This annotation of information allows for items to be categorized and labeled, thereby facilitating operations such as navigation, filtering, and searching [9].

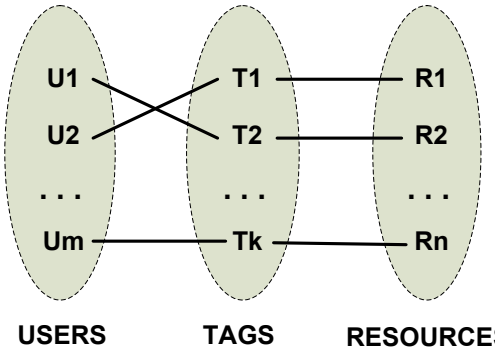


Figure 1: Tags acting as an intermediary through which users can discover resources[1].

Tags have been observed to perform a variety of different roles in categorization, and are able to, amongst other things, ascribe resources with specific attributes. According to Golder and Huberman [9], tags are often associated with one of the following functions:

- Identifying who or what the resource is about through the utilization of common nouns.
  E.g.: *sport, politics*.

- Identify what the resource is.
  E.g.: *article, book.*

- Identifying the resource owner.
  E.g.: *Tom, mine*.

- Identifying qualities or characteristics, largely through the use of adjectives.
  E.g.: *silly, funny, upsetting*.

- Self Reference, identifying content in terms of its relation to the tagger. E.g.: *mine, mypics*.

[1] Adapted from diagram in [8]

- Task Organizing, by attaching tags to particular tasks.
  E.g.: *toread, todo*.

Tags are therefore able to effectively impart a multitude of different metadata to electronic resources

2.2.2 Folksonomies

Folksonomy is a term that is often used to describe the practice of collaborative tagging. It was conceived by Thomas Vander Wal [6] to describe the growing tend of collaborative classification. While the term is a combination of the words *folk* and *taxonomy*, folksonomies fail to exhibit the hierarchical structure that the latter are characterized by.

Although tagging is by no means a novel exercise, websites such as flickr[2], and most notably del.icio.us[3], have inspired ordinary Internet users to collaborate in these shared spaces by tagging content, and browsing the content tagged by others [8]. These practices have highlighted several advantages and disadvantages that can be attached to folksonomies in general, and are discussed below.

- *Advantages*

Most notably, folksonomies are inclusive in the sense that they incorporate the thoughts of all involved, regardless of any social or political predisposition. Individuals are not restricted by controlled vocabularies that enforce a predefined structure, but are rather empowered to express their personal needs as they please. As such, folksonomies exhibit a contemporary representation of trends that static structures are unable to replicate. This adaptability enables folksonomies to overcome the regimented classification schemes inherent in hierarchical structures. Items are able to exist dynamically in the midst of other items, instead of being forcefully classified according to predefined relationships. As a consequence, folksonomies effectively facilitate resource discovery, allowing for unexpected or even unknown items to be easily and democratically incorporated into a self-moderating structure. A Folksonomy is therefore an expression of the direct information needs and desires of its users [11].

Folksonomies present additional benefits in terms of cost and usability when compared to taxonomies and ontologies. While the latter are difficult and costly to construct and maintain, folksonomies place the onus on the user to attach meaning and semantics to resources [8]. These exercises are assisted by the accessibility and ease of use that have become associated with folksonomies, and are aided further by their participative nature.

[2] http://www.flickr.com
[3] http://del.icio.us

- *Disadvantages*

While the benefits of folksonomies are certainly noticeable, much debate continues to surround their failings. The most significant shortcoming of these structures is their lack of hierarchical organization. Tags, therefore, exhibit no intrinsic relationship between one another [6], and possess no means to explicitly express their significance to other tags. Although humans are able to easily establish the relationship between tags and the semantics that they represent, machines fail to make these implicit connections because of their explicit nature. As such, folksonomies remain largely as latent stores of data for machines given their limited reasoning abilities, thereby rendering most of their content as strictly human-friendly.

In addition to the abovementioned concerns are the problems of duplication and imprecision. Because collaborative tagging occurs in largely unsupervised environments, there exist few mechanisms to control the occurrence of polysemy and synonymy. Select words have several related meanings associated with them, and often are instrumental in deteriorating the quality of search results through their association with other inapplicable items [9]. Furthermore, synonymy, where several words have the same meaning, represents another point of aggravation. While users often interpret resources in comparable ways, the terms that they tag these resources with are often dissimilar. *Motorcar*, *vehicle*, and *car*, are all synonymous terms that can be assigned to the same resource. However, a query using the term *car* will only return resources that match that tag, and will often fail to recognize its association with *motorcar* and *vehicle*. Folksonomies, therefore, do not readily facilitate targeted searching [4]. Another symptom of the abovementioned difficulties is the rate at which the number of tags that are associated with a system grows [9]. The number of tags that are present within particular systems have been noticed to increase dramatically over time.

Unfortunately, uncontrolled, shared environments often lend themselves to misuse, as is sometimes the case with folksonomies. Unethical users often associate resources with blatantly inappropriate tags in an attempt to corrupt the system. Although this nevertheless poses a problem, collaborative efforts such as Wikipedia[4] have shown that such uncontrolled endeavors are indeed possible.

While misuse is definitely a concern, problems often arise out of disagreements alone. Tagging is subject to conflict between contributors, where differing opinions and perspectives can lead to disagreements and potential power struggles [9]. As such, larger tagging systems often consist of distinctly personal categories that exist alongside ones that are widely agreed upon [9].

[4] http://www.wikipedia.org

2.2.3 Tag Clouds

Tag clouds are visually weighted renditions of collections of terms that can be used to represent and interface to the concepts that are contained within collections of information [12]. They are generally envisioned as a collection of user-defined tags that can differ in visual appearance, depending on their prominence. Their somewhat graphical nature and ease of use have contributed to their popularity as a method to support and facilitate the navigation and retrieval of tagged data. As such, tag clouds act as a window through which the prevalent concepts of a folksonomy can be viewed, and can therefore be interpreted as a stigmergic structure that provides an aggregate of tag-usage statistics [13].

Figure 2: A tag cloud, ordered alphabetically. [5]

Rivadeneira (et al.) [14] suggest that the appearance of tag clouds is often a function of two factors:

- Text Features
  - *Font Weight*: the weight or bolding of text to represent frequency of an underlying structure. Generally, the greater the weight, the more prominent the item.
  - *Font Size*: the size of the font is analogue to popularity of the term that it is applied to. Generally, the greater the size, the more prominent the item.
  - *Font Colour*: font colour is utilized in a variety of ways, conveying importance or relevance by varying the appearance of the word.
- Word Placement
  - *Sorting*: terms within the tag cloud can be presented alphabetically, by frequency, or even arbitrarily.
  - *Clustering*: words can be sorted semantically, or the user can specify their preference.
  - *Spatial Layout*: words can be presented in sequential lines, or in more sporadic layouts.

[5] http://blogs.talis.com/nodalities/gatescescloud.jpg

By the utilization of combinations of the abovementioned factors, tag clouds are able to facilitate the following tasks [14]:

- Search: tags allow users to efficiently locate specific terms, and often act as a means to access underlying content. Specific keyword searches, however, are often ineffective.

- Browsing: users are able to easily browse a website's principle content.

- Impression Formation and Gisting: tag clouds are tools that empower users to easily formulate a general impression of the underlying content that is associated with a domain.

- Recognition/Matching: users are able to easily determine whether a particular data set represents their particular interests. Tag clouds are able to offer a glimpse into the types of resources that are present within a particular domain.

2.2.4 Stigmergy

According to Wikipedia, Stigmergy is a method of communication in decentralized systems in which the individual parts of the system communicate with one another by modifying their local environment [15]. The term was proposed during the 1950's by Pierre-Paul Grasse'[16] to define the usage of environmental modifications to permit forms of indirect communication between individuals in social insect societies.

Although the concepts of Stigmergy exist primarily in the realm of the natural sciences, the phenomena that are expressed by it have been observed in technologies such as the Internet [17]. Blogs, news, forums, and more recently, tag clouds, have illustrated qualities that present resources to users that exhibit a visible form of social weighting. In a similar way to how ants invest items with their pheromones to signify importance, users of tagging systems impart their opinions by selecting particular items within a tag cloud. While pheromone-intense areas have a tendency to attract the attention of other ants, popular tags that appear larger and have greater visual appeal have been shown to be more noticeable to users. In both situations, stigmergy is personified as a form of passive influence.

Furthermore, as pheromones fade over time due to transforming environments, trends in user perception follow similar patterns in the light comparable circumstances. In this sense, stigmergy offers a contemporary view of the world towards which it is focused.

Apart from these scenarios existing as an interesting correlation, their likeness suggests that further stigmergic connections between the natural and digital world can be made through investigation relating to folksonomies and tag clouds.

3. SPECIFICATION FOR SOLUTION

The primary intention of the project around which this report is based is to investigate novel mechanisms that are able to create data that is present and current. Although managed hierarchical structures, such as taxonomies and ontologies, are useful mechanisms of conceptual representation, they are inherently static in their ability to adapt to a perpetually changing user base. Current political and social events continually change the associations that individuals place between these concepts, rendering current data models as stagnant. The proposed system will attempt to combine both manual and automatic methods of categorization in an attempt to provide semantic data that is both current and useful.

3.1 The Manual Tagging System [Ian Saunder]

A tag, as it relates to information technology, is generally a keyword that is associated with or assigned to a resource. Tags are often grouped together into a collectively shared area called a tag cloud. This is static in nature and shows each tag's frequency by changing its appearance. By extending this system to emphasize the importance of spatial relationships amongst these tags, it is envisioned that associations between these tags can be measured and converted into a useful form.

Figure 3: An example of a populated tag cloud that emphasizes spatial importance

A system with the abovementioned qualities was implemented, and required its users to read a piece of text which they would then tag with a term that appeared within it. Users were then able to reposition these tags so that related tags were in close proximity to one another while at the same time being distant to ones that they had little relation to.

3.2 The Automatic Document Indexing System (using Probabilistic Latent Semantic Analysis) [Sean Colledge]

To contrast the manual tagging system, an automatic document indexing algorithm was implemented by the author's project partner. The algorithm, known as Probabilistic Latent Semantic Analysis (PLSA), has become a popular means of ascertaining the topic of the text upon which it is processed.

3.3 Interconnection

These two separate components that constitute the project as a whole will share identical inputs. As such, their respective outputs will be compared and possibly combined to form

data that will hopefully provide a meaningful insight into the collective understanding of this restricted vocabulary.

4. DESIGN [Refer to Appendix A]

4.1 High-Level System Description

A successful solution is likely to comprise several distinct, modular components that each contribute to the overall functionality of the system in a clearly defined way. A high-level analysis suggests the inclusion of three distinct entities: the *browser component*, the *web server*, and the *database server*.

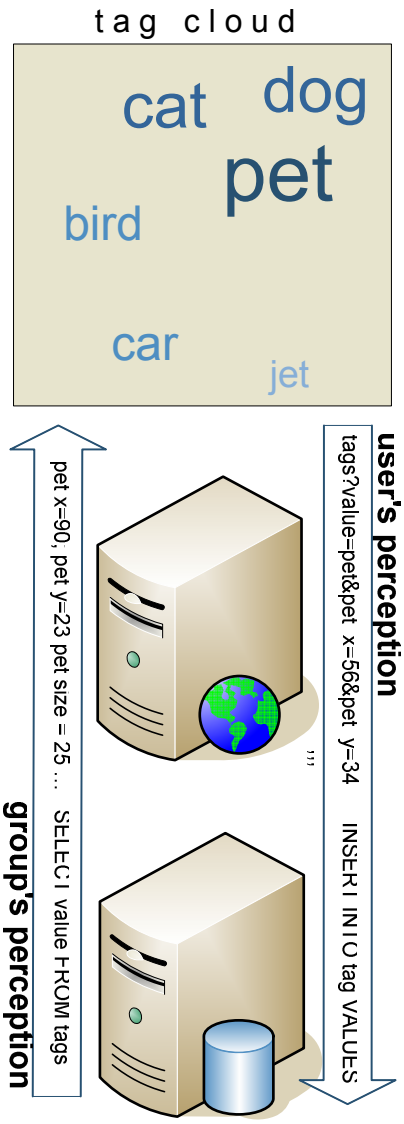


Figure 4: A high level system diagram showing interconnections and flow.

4.1.1 The Browser Component

This module represents a focused area of interaction between the system as a whole and the users that enter the data from which inferences are finally drawn. Written in Perl and a combination of HTML and JavaScript, the module retrieves tag information from a web server before translating this information into a tag cloud which offers greater visual appeal and interaction. Users are then able to enter tags and spatially manipulate them, along with the tags that were entered previously and reloaded by default. As such, this component is run remotely on the clients' machine. Once all tagging activities have been completed by the user, this data is returned to the web server, which then initiates the process of permanent information storage.

4.1.2 The Web Server

There exists a significant connection between the Browser component and the Web Server component, and is consistent with the traditional Client-Server model. Requests issued by the browser component, on behalf of the user, are serviced by the Web Server by responding appropriately to the given request.

4.1.3 The Database Server

The database server is responsible for managing the database that facilitates the permanent store of information relating to all modules (including both the tag and PLSA modules). Data is retrieved from the database by the browser component in a manner that

is consistent with a user's request, and then returns inputs to the database for permanent storage.

MySQL is to be used as the database because of its price (*free*), its availability (*installed on the departments servers to which we have access*), and its familiarity (*we've used MySQL many times before*).

4.2 High-Level System Interaction

4.2.1 Load

i. A load action is triggered by an event in the Browser Component, and forms part of any action that requires the retrieval of data from the database. The action in question is a SQL query that is embedded within a Perl program. The process is initiated by a client's request for such a file.

ii. The web server then executes the requested file, and performs any embedded database queries.

iii. The database server executes the query. For a load, only SELECT statements form part of the query.

iv. The database server then returns a table of results to the web server.

v. Having received the results of the query from the database server, the web server creates an HTML response, including the results from the database query wherever needed.

vi. The web server then sends the generated page to the requesting client.

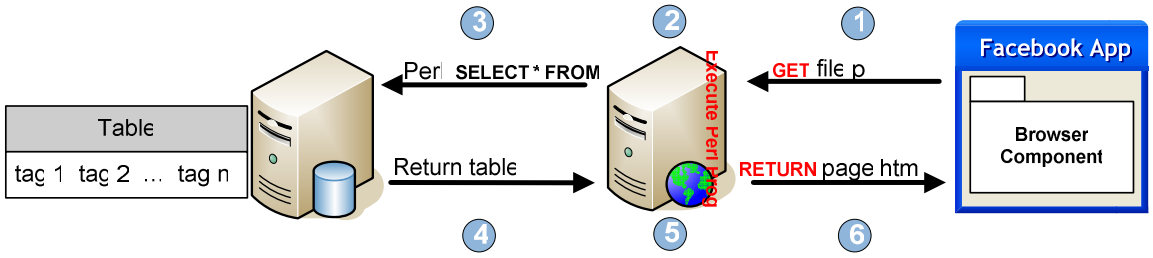


Figure 5: A system diagram depicting a Load action

4.2.2 Save

Steps i through iii are identical to those executed in a *Load*. However, there is no need to generate an HTML response, making the remaining steps redundant.

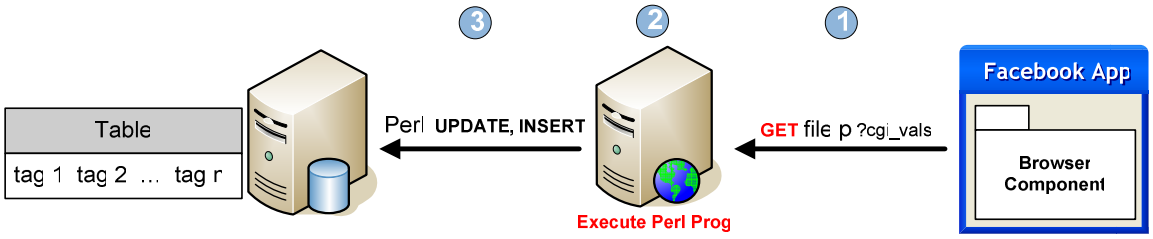


Figure 6: A system diagram depicting a Save action

4.3 Database Design [See Appendix A for Entity Relationship Diagram]

4.3.1 User Anonymity

By associating a tag with a particular resource, the acting user imparts a sense of understanding specific to them that is chosen under the assumption of anonymity. As such, it is the supposition of most users that little, if any, connection can be made between them and their tagging practices. A moral obligation, therefore, is placed upon the developer to ensure that the users' privacy remains as confidential as possible while still allowing for effective system operation and data gathering.

The abovementioned consideration necessitated special concern relating to the design of the database, and dictated that connections could not be made between a user and the tags that they create. As such, the database was designed in such a way that associations that could undermine the users' privacy could not be made.

4.3.2 Tags

In addition to recording a tag's position, its center co-ordinates are preserved. These values are a function of the tags top-left co-ordinates and its size, and are used when computing distances between tags. A field exists to store whether a tag's position had been changed between tag cloud sequences, as well as whether that particular tag was the term chosen by the user during the tagging of an article.

4.3.3 Tag Cloud State

Instead of only storing the current state of the tag cloud, it was decided to record each stage of the cloud's development. Although this is not necessary for successful operation, it provides data which depicts the progression of each tag cloud from its inception onwards. As such, data exists to allow for a visualization of the tag cloud's metamorphosis, and may prove for an interesting addition should time permit.

4.4 Algorithms

Because tag clouds are weighted representations of a collective understanding, an algorithm was developed to produce a similarly reflective representation of the users'

perception of spatial associations. After having manipulated the tags' positions, it is essential to consider these changes in proportion to the number of preceding changes in arrangement.

4.4.1 Weighted Tag Movement

After careful consideration, a decision was made to weight the movements of a tag in proportion to the number of times that its position had been previously altered. For example: if a tag had been repositioned on three previous occasions, it will only move ¼ of the distance expressed by the user new relocation. Therefore, tags which are not moved do not increasingly oppose change as others are moved around them.

```
consider tag's old and proposed positions

  find change in X co-ordinates
  find change in Y co-ordinates

    weighting = 1 / (number of previous movements + 1)

    tags new X = change in X * weighting
    tags new Y = change in Y * weighting
```

Figure 7: The weighted tag movement algorithm

4.4.2 Algorithm Weaknesses

Although the abovementioned algorithm appears to adequately provide a means of weighted collective input, it is subject to forms of unintended use.
An assumption that the algorithm makes is that the input is an accurate reflection of the tag cloud that the user deems most appropriate. An understanding of the weighted tag placement algorithm equips the tagger with the ability to increase the influence of their inputs. Because tags are moved a fraction of the users desired change in position, it is possible to intentionally reposition a tag further than would be necessary or appropriate. If, for example, the user wished to reposition a tag from its present location of *x1, y1* to a new position of *x2, y2*, the user could purposely alter the proposed location by dragging it to *x2 + r, y2 + r*, thereby increasing the weighted change in position. While an understanding of the algorithm is required to perform these manipulations, it is nevertheless a possible weakness.

5. IMPLEMENTATION

The emphasis placed upon data collection ensured that the browser component featured as the principal element of the system. While the web server and database components were a necessary inclusion to the system, they existed primarily as a means to service needs that the browser component itself could not fulfill.

As had been previously proposed, the system was comprised of three key components: the browser, web server, and database server components.

5.1 The Browser Component

5.1.1 The Application Platform

After much deliberation and uncertainty, it was decided that the browser component was to take the form of a Facebook application. Facebook is a popular social-networking website that is popular amongst students. Several reasons supported our decision to use Facebook as a means to gather data:

i. Facebook is a well-liked and popular means of digital interaction amongst people and groups around the world. The addition of custom applications to the website, which allows account holders to create unique applications that can be added to the profile of Facebook users, has ensured its continued growth and success. As such, several support mechanisms are available to application creators, and are well maintained by a network of fellow developers.

ii. The author's extensive social network within Facebook ensured that a large number of potential users were immediately accessible through the many communication mechanisms provided by Facebook. Hundreds of users could be notified of the application existence within a single hour, making Facebook an appealing platform.

iii. The author, having developed a Facebook application for a completed module, has experience and prior exposure to the Facebook API.

5.1.2 Data Sources and Content

A tag, as it relates to information technology, is generally a keyword that is associated with or assigned to a resource. It therefore comes as little surprise that specific resources have to exist in order for tagging to take place. Such a resource would have the following attributes associated with it:

i. The resource must be text-based. This is a necessity as the PLSA algorithm needs to be applied to the resource.

ii. The resource should be freely available and modifiable to suit the purposes of the application. This requirement ensured that several worthy sources had to be disregarded because of the stipulations detailed within the copyright.

iii. The resource must be appealing, and not laborious for the user to read or analyze. Humorous and entertaining texts would aide in creating a stimulating environment that would hopefully encourage the user to take pleasure in and appreciate the tagging process.

UnNews is a humorous and often completely inaccurate news source that forms part of Uncyclopedia, the content-free encyclopedia. The latter exists as a parody to Wikipedia, and has seen dramatic increases in popularity over recent months. UnNews is freely available under a Creative Commons license which allows for its use and modification, making it ideally suited to the purposes of the application. The application content, therefore, exists in the form of specific news stories taken from UnNews which are humorous and not offensive.

5.1.3 Application Structure

The limitations of the chosen development languages posed particular limitations to the system design, and were ultimately responsible for the applications final internal structure. Of primary concern was JavaScript's inability to perform database accesses, and Perl's server-side predisposition. Once a web page event had been triggered, there was no possible means of immediate database access. As such, relevant data had to be transferred to another Perl file by means of CGI, and would then be committed to the database. Therefore, any application page that needed to insert into or update existing database tables relied upon an accompanying *db_writer* file. A POST was used to transfer the data, owing to its increased security and capacity over GET. These are mechanisms for parameter passing over HTTP.

5.1.4 Application Layout

i. Home

The Home page is loaded when the user loads the application. It contains a description of the application, instructions for mystified users, and statistics relating to recent activity.

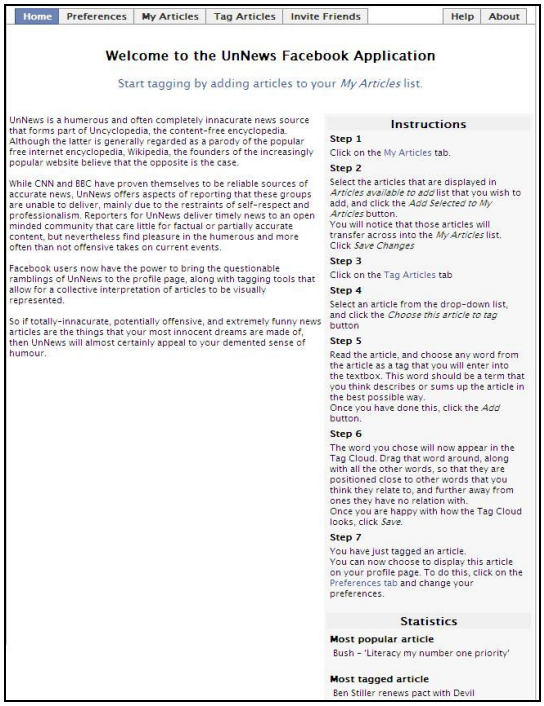


Figure 8: The UnNews Home Page

ii. Preferences

This page allows the user to choose a desired method of selecting the article that is to be displayed on their profile page. Three options allow the user to either display a random article, the most recent article, or a specific article. Users are only able to consider articles that form part of the *My Articles* list.

iii. My Articles

Users are able to select articles that they have interest in by adding them to their *My Articles* list. This list exists to facilitate the *Preferences* function, as this section only considers articles that form part of it.

iv. Tag Articles

This complex page enables users to tag and manipulate the tag clouds of selected articles. After initially selecting an article to tag, the article text is show alongside its most current tag cloud. Once a tag has been entered, the user is able to change the position of all tags within the cloud to better suit their interpretation of the relationships between the terms.

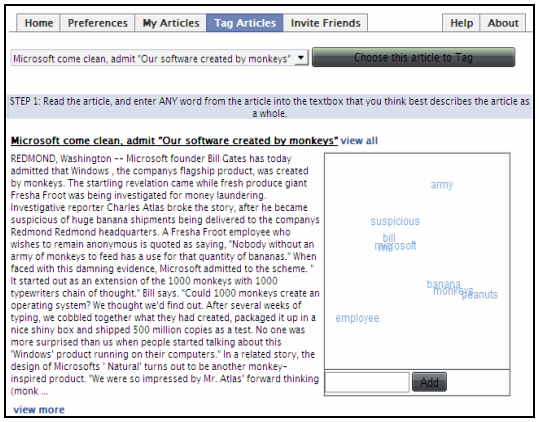


Figure 9: The UnNews Tag Articles Page

v. Invite Friends

The Invite Friends page provides a useful mechanism that allows users to easily notify their friends of the application.

vi. Help

The help page explains the purpose of the application, and provides answers to frequently asked questions.

vii. About

5.1.5 Tag Cloud Appearance

While tag clouds are often merely collections of terms that are listed in some predefined order, their arrangement is made more noticeable and distinctive by the varied appearance of the tags from which they comprise. These user-entered terms are dissimilar in their size and colour, with both of these qualities being directly related to the tags prominence. Although the tag clouds that appear in the UnNews application have the added dimension of spatiality, the tags that form part of the cloud are nevertheless distinguished in a similar way.

The font size and colour of a tag is directly related to the number of times that it has been selected by the user and entered into the tag cloud. As this number increases, the size of the tag increases and it adopts a darker shade of colour.

Font Size
The relationship between the number of times that a tag has been added and its size is best described as a logarithmic function. A linear function, where the font size would increase in direct proportion to its frequency, would be unsuitable as popular tags would adopt sizes that would quickly render them unable to fit within the dimensional constraints of the tag cloud. In contrast, a logarithmic function allows for the size of a tag to initially increase quickly, and then decrease in its rate of growth as its popularity continues to grow.

$$font\ size = ln\ (\ frequency\ )\ *\ constant$$

Figure 10: The formula used to determine tag font size

- *The Constant Factor*
  A constant factor was used to scale the font size in order to render it visible. Although the logarithmic function ensures a favorable range of font size, simply setting the fonts' size to this value would result in tags that are completely unnoticeable. The constant factor remedies this problem, while at the same time ensuring that size the rate of increase is continually declining. Trial and error proved '8' to be a suitable constant value, as values above and below it seemed to either result in abrupt or painfully gradual increases in font size.

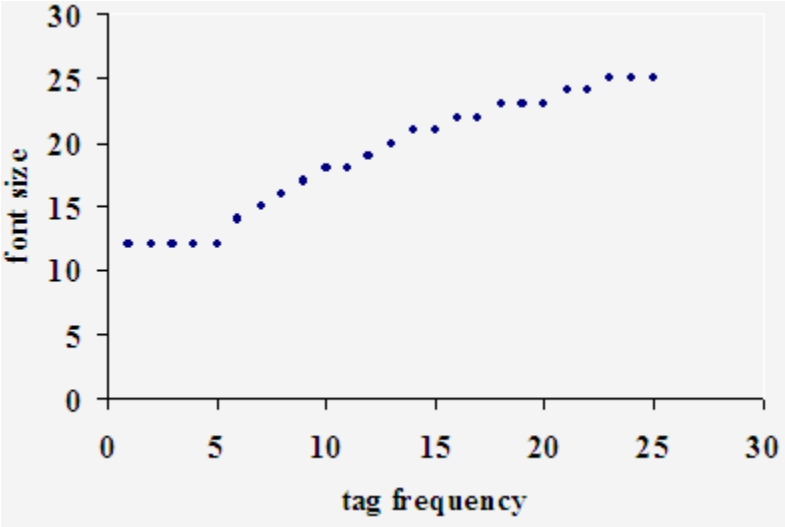


Figure 11: A graph showing font size against tag frequency

- *Default Font Size*
  When frequencies less than 5 are applied to the above formula, font sizes are produced which are too small to be visible. As such, a tag is given an initial size of 12 points, and is incremented should the above formula produce an output which is greater than this value.

Font Colour

The font colour of a tag also changes, depending on the tag's frequency. More specifically, the more frequent the tag, the darker its font colour becomes. This change in appearance is related to the change in the tag's size, with both of them being altered at the same time. As such, tags do not change size without changing colour as well.

Font Colour	Font Size (<=)
	12
	14
	15
	16
	17
	18
	19
	20
	21
	otherwise

Figure 12: A table showing the relationship between font colour and font size

5.1.6 Application Data

The UnNews website hosts hundreds of articles that have accumulated over the last two years. However, the number of articles that formed part of the application had to be limited to ensure that tagging events were not spread too thinly. Adding dozens of articles would likely result in sparse and inadequate tag clouds. As such, the number of articles available for tagging was limited to eleven for the entire duration of the data gathering process.

5.1.7 The Tagging Process

In order for a tag to be created and associated with an article, several actions have to take place. The process is initiated by the user loading the *Tag Articles* page, and selecting an article to tag. After having selected an article, the article text is displayed alongside its present tag cloud. An instruction is presented above the article and cloud informing the user to read the article and select a single word from it that they think provides the best synopsis of the text. The limitation of being able to only use words from the article itself stems from the necessity to execute the PLSA algorithm on the same data that is presented to the user. This ensures that the outputs are bounded by an identical range of inputs, allowing for the results of the tagging and PLSA algorithms to be compared. As such, inputs were limited in that character sequences could not be entered that did not correlate to those of words found within the article text.

Tag cloud dimensions of 230 pixels square were decided upon, as this seemed to provide a large enough area around which tags could be dragged, while at the same time remaining visually appealing and unobtrusive.

Because of the limitation imposed on tag choice, mechanisms were introduced to ensure that tag discovery and selection were as straightforward as possible. A predictive text feature was added to the tag cloud to aide in this regard, and displayed a list of words from the article whose earliest letters matched those that match the input of the user. As further characters are entered, the list shortens to correlate to the revised input. Clicking on an item in the predictive text list will add the entire word to the text box. Deleting characters from this input would increase the size of the predictive text list.

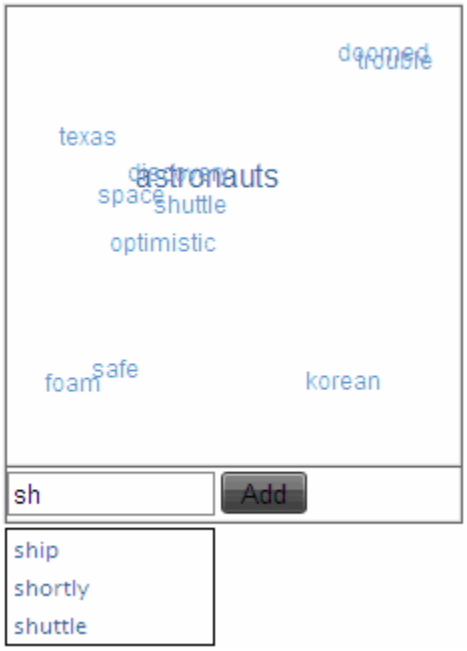


Figure 13: A tag cloud showing the predictive text feature

After having entered a term, a further instruction is displayed informing the user to reposition the tags so that related terms are close together while terms with little relation are far apart. The instruction is intentionally presented as being open ended as spatial interpretations and relationships are uniquely personal and cannot be easily quantified. The purpose of tagging is to impart an element of individually-inspired data, and as such, is a function of the user.

5.2 The Web Server

The web server is responsible for appropriately responding to and servicing the requests of the client made via the browser component. The services of a pre-existing departmental server were acquired with the kind permission of Dr. Hussein Suleman. Banzai, the apache web server in question, had had the Facebook API pre-installed as it

was used to host the Facebook applications developed for a Computer Science Honours course. Banzai, therefore, was ideal in that its ability to successfully host Facebook applications had been previously confirmed.
No software modifications were made to Banzai, and the web server's configuration was left unaltered.

5.3 The Database Server

Casper, a departmental server, was used as the application's database server. My SQL 4.1.7 was installed on the machine, and was responsible for all permanent storage relating to the Facebook application.

6. DATA PRESENTATION [See Appendix B]

6.1 Method of Presentation

The final tag cloud states are reproduced and presented as Appendix A. Accompanying these images are statistics relating to the tags that were entered into the clouds, and shows each tag's frequency and its final position.

6.2 Statistics

A total of eleven articles were selected as a source from which users could select terms to tag with. After ten days of activity, 310 tags were created by the various users that added the UnNews Facebook application.

6.3 Article Reference

From this point onwards, the articles that the system utilized will be referred to using an integer reference as is described below.

Article ID	Article Title	Article URL http://uncyclopedia.org/wiki/ UnNews:
1	F1 'ace' Alonso accepts defeat gracefully	"ITS_NOT_FAIR_I'M_THE_BEST_NOT_HIM_"-_F1_'ace'_Alonso_acedes_defeat_gracefully
2	Ben Stiller renews pact with Devil	Ben_Stiller_renews_pact_with_Devil
3	A nice cup of tea and a sit down proven to cure all human ills	A_nice_cup_of_tea_and_a_sit_down_proven_to_cure_all_human_ills
4	13 year old wins spelling bee; guarantees life of virginity	13_year_old_wins_spelling_bee%3B_guarantees_life_of_virginity
5	"Facebook stole my face" student claims	"Facebook_stole_my_face"_student_claims
6	Bush – 'Literacy my number one priority'	Bush_-_'Literacy_my_number_one_priority
7	UK floods turn out to be advertisement for "Evan Almighty"	UK_floods_turn_out_to_be_advertisement_for_Evan_Almighty
8	Doomed Astronauts Optimistic	Doomed_Astronauts_Optimistic
9	Gas stations begin giving away free gasoline	Gas_stations_begin_giving_away_free_gasoline
10	Mr. Potato Head busted	Mr._Potato_Head_busted
11	Microsoft come clean, admit "Our software created by monkeys"	http://uncyclopedia.org/index.php?title=UnNews:Microsoft_come_clean_admit_Our_software_created_by_monkeys

Figure 14: A table showing the articles used by the UnNews Facebook Application with accompanying reference ID's

7. DATA INTERPRETATION

7.1 Tag Cloud Representation

Several features of tag clouds, most of which have been previously mentioned, are inherent in all of their implementations. Attributes such as font size, weight, and colour are all qualities that help to distinguish elements that constitute the abovementioned assembly of words. An extension of the tagging metaphor, however, ensures that extensions to these attributes should be made simultaneously.

The extended tagging practice proposed and presented within this report introduces the concept of spatial significance and usefulness, and allows for tags to be repositioned within the tag cloud that they reside. As such, additional attributes are required to be associated with each tag so that these augmentations can be recorded. Also to be noted is the fact that a tag's colour is a function of its size, and is therefore not recorded.

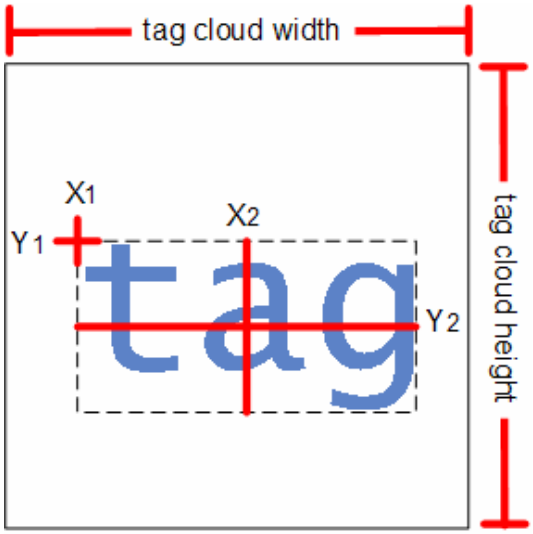


Figure 15: A tag cloud with its center and top-left coordinate markings shown

7.1.1 Defining a Tag's Position

In addition to storing a tag's font size, its relative position within a 2-dimensional Cartesian plane is recorded, as well as the tag's center position. The positions are relative in the sense that they are evaluated in relation to the confines of the tag cloud itself and not merely recorded as its absolute screen position. To facilitate this calculation, each tag is created within a tight-fitting *DIV* HTML element that snaps to the outermost points of the tag text, and allows for screen positions to be easily obtained.

The following equations describe the calculations required to compute the center points of a tag:

$$\textit{center-x co-ordinate} = \textit{left-x co-ordinate} + (\textit{tag width} / 2)$$

$$\textit{center-y co-ordinate} = \textit{top-x co-ordinate} + (\textit{tag height} / 2)$$

Figure 16: Formulas used to calculate a tag's center coordinates

7.1.2 Tag Movement

Every tag is able to be repositioned using a drag and drop interface. A tags movement is restricted to the dimensions of the tag cloud within which it resides, ensuring that they are

unable to move outside this area. The outermost corners of each tag provide a means to prohibit its movement beyond the edges of the cloud.

7.1.3 Tag Cloud Dimensions

Specific dimensions were decided upon for all tag clouds that were created. This was to ensure that the data produced by all clouds were immediately correlated in scale, as well as to provide a consistent means of interaction for all users. The tag clouds are all square in shape with sides measuring 230 pixels each. These dimensions proved to be great enough to allow for tags to be freely repositioned without constriction while at the same time ensuring that the clouds' size was not a hindrance to surround page elements.

7.2 Measuring Euclidian Distance

Euclidian distance is analogue to the distance obtained by manual measurement using a ruler or similar device. This measure of distance is a sensible choice in that it provides a simple means of interpreting relative differences in spatial orientation. [18]

$$d = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

Figure 16: The standard Euclidian distance formula

The equation presented above is the general form of a 2d specific Euclidian distance formula. An important observation is that the center co-ordinates of a tag are considered when using the distance equation. The reason for this is that font size is not uniform through out the tag cloud as tags with higher frequencies appear larger than those that are less frequent. As such, the top left co-ordinates of a tag become less influential or descriptive of its orientation.

To resolve this issue, the center co-ordinates of the tag are considered, as is shown in the image below.

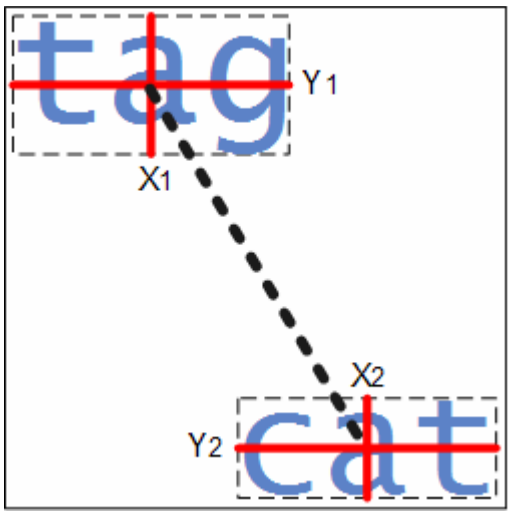


Figure 17: Two tags with center coordinate markings that connect a distance line between them

7.3 Converting Euclidian Distances to Probabilistic Relations

While the addition of spatial significance to tag clouds ensures that further pertinent relationships between tags can be observed, the data inherent in these associations nevertheless remains latent until it can be transformed into a more useful, non-specific form. For instance, a statistic showing that a pair of tags is separated by a certain distance is of use to no one. For this reason, it is of primary concern to transform Cartesian distances into probabilistic relationships, making these relationships more useful and accessible.

7.3.1 Correlations between Euclidian Distance and Probabilistic Relations

Before exact values can be calculated that detail the relationship between the abovementioned measurements, the nature of correlation between them has to be affirmed. It is a natural intuition to group items, regardless of their form or nature, according to their similarity and relevance to one another. Elements that are comparable and related, therefore, are placed together, while those that are not are separated in some way. For this reason, instructions were given to the users of the Facebook application informing them to place related tags close to one another while at the same time ensuring that unrelated ones are separated. Therefore, as the space that separates a pair of tags increases, their probability of relation decreases in a certain fashion. This exemplifies a negative correlation between the magnitude of separation and the probability of connection or association.

7.3.2 Functions which do not describe the correlation between Euclidian distance and probabilistic relation.

It was initially the opinion of the author that an inverse linear relationship exists between the Euclidian distance that separates two tags and their associated probability of relation. Such a relationship could be exemplified by the following equation:

$$probability = \frac{\alpha}{distance}$$

α is a constant of proportionality

Figure 18: A basic linear equation relating probability to separation distance

After having consulted with lecturers in the department, it became increasingly difficult to justify the existence of a linear relationship between displacement and probability. Although these were sentiments that the author shared, they could not be supported at the time with experimental data. As such, a small test group was assembled and a simple test undertaken in an attempt to better understand a user's perceived relationship between distance and strength of connection.

Determining how far unrelated tags are placed from one another

While overlapping tags were interpreted as having a definite connection, the correlation between the users' placement of unrelated tags and their spatial interpretation of 'unrelated' were unobvious. These individual and therefore totally unique interpretations ensured that users may interpret tags as being totally unrelated, but will nevertheless position them at different distances to one another. As such, there exists no singular and absolute difference in spatial displacement between two tags that indicates that they are totally unrelated to one another. While one user may separate two unrelated tags by the maximum distance permitted by the cloud (*its hypotenuse*), another may use half that distance to describe an identical point of view regarding their connection.

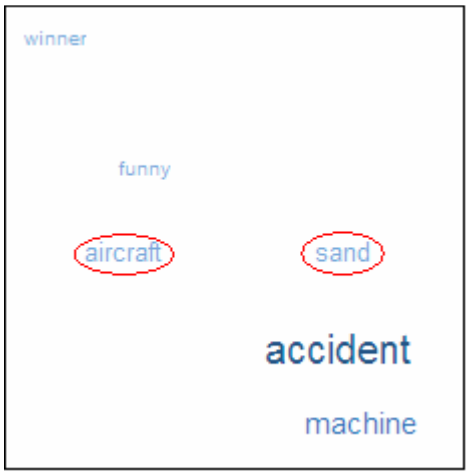


Figure 19: The tag cloud used during the experiment, with focal tags shown within a red oval.

Participants were asked to consider the tag cloud above and instructed to decide for themselves the strength of the connection between the tags *aircraft* and *sand*. These terms were chose as they are noticeably unrelated to one another. After having assessed the strength of association, the test subjects were asked to reposition the two tags according to the principle that related tags should be close to each other while unrelated tags should be far apart. The final positions of the tags were recorded and the distance between them computed using the Euclidian distance formula.

The results of the test show an average displacement of 212 units (in this case pixels), suggesting that a large area of the tag cloud is utilized during repositioning. It is reasonable to presume that humans are unable to consistently judge distances to a degree of high accuracy, especially when those distances belong to a small scale. If only a small space was used by users to express unconnected relationship, users might find it easier to be consistent in their placement of other tags, given the assumption that smaller distances are possibly easier to consider.

	chosen separation
user 1	228
user 2	199
user 3	168
user 4	263
user 5	237
user 6	234
user 7	188
user 8	148
user 9	198
user 10	254
	212 average

Figure 20: The chosen separations of all test participants

It would appear that tags are dragged to positions where they appear to appropriately represent their relation to other surrounding and distant tags. This *sense* of grouping suggests that a linear function does not accurately represent the correlation between distance and probability, especially when one considers that a large part of the tag cloud is used to express the former. The idea that unrelated

tags are placed at distances which seem *far enough* from each other suggests a more flexible relationship.

7.3.3 An Inverse Exponential Decay Function

The unsuitable nature of a linear function depicting the relationship between the distance of separation and associated probabilistic relation ensures the need for further investigation. An inverse relationship, however, most certainly exists between these two quantities, irrespective of the nature of the function describing them. Candidate solutions include the logarithmic and exponential functions, owing to their favorable slope. While the coefficients of both can be manipulated to form similar curves, the exponential decay function is naturally bounded between the y co-ordinate values of 0 and 1, thereby making it immediately suitable for probabilistic application. The function is bound in such a way because N_0 is taken to be 1.

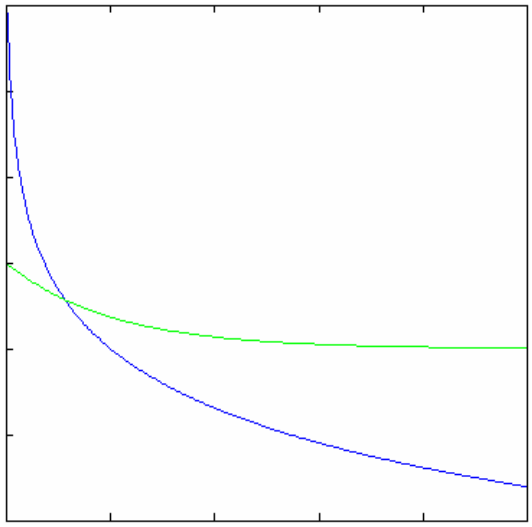


Figure 21: A plot showing exponential (green) and logarithmic (blue) functions

For this reason, the exponential decay function was chosen to model the relationship between the distance of separation and associated probability between tag pairs.

Function Form

The standard form of an inverted exponential decay function is presented below, along with an explanation of its symbolic representation:

$$N(t) = N_0 \, e^{-\lambda t}$$

Symbol	Description	Value
$N(t)$	*function value at time t*	(0, ∞)
N_0	*function value at t = 0*	(0, ∞)
e	*base of the natural logarithm*	2.7183
λ	*rate of decay*	(0, 1)
t	*time*	(0, ∞)

Figure 22: The basic form of the Exponential Decay Function, accompanied by symbol explanations

An examination of the $e^{-\lambda t}$ component demonstrates the function's horizontal bounding between 0 and 1. Although the function is continuous, only positive values of t are considered. This is because distances are regarded as scalar quantities and do not convey any sense of direction as would be implied by a vector quantity such as displacement. In addition to this, the fact that the rate of decay, λ, is always a positive number ensures that the exponent of the term e is always a negative value. As such, $e^{-\lambda t}$ will always evaluate to a positive number that is greater than 0 or less than or equal to 1.

Tag Specific Functions

Although not supported by experimental data or proven as fact, it is the opinion of the author that the function that relates spatial separation to associated probability is unique for each tag cloud. As such, the characteristics of a tag cloud are certain to have an affect on the tagging activities of the users that manipulate them. Factors that support this notion are discussed below:

- *Number of tags*
  It would appear that as the number of tags within a tag cloud increased, the spatial displacement required to show strong relations between them would decrease. Tags would have to be closer to other tags that they related to in order to be certain that the association is not confused with other surrounding tags.

- *Tag Clustering*
  The concentration of tags within the tag cloud would suggest that the required distance between tags to show their connection would decrease. Tags would have to be placed closer to tags that they share a strong connection with to ensure that this relation is not inadvertently confused as a positive correlation to a nearby tag of less relevance.

Because of the abovementioned considerations relating to tag cloud uniqueness, the coefficients of the exponential decay function were altered to better represent the spatial data extracted from the tag clouds. As such, a function of the form presented below was considered:

$$P(d) = P_0 \, e^{-\lambda d}$$

Symbol	Description	Value
$P(d)$	*function value at distance d*	(0, 1)
P_0	*function value at d = 0*	1
e	*base of the natural logarithm*	2.7183
λ	*1 / average separation*	(0, 1)
d	*separating distance*	(0, 310)

Figure 23: The modified Exponential Decay Function, accompanied by symbol explanations

The above explanation illustrates two significant changes to the values of the function coefficients:

- **$P_0 = 1$**
  The value of the function when the distance separating two tags is zero is equal to 1. This implies that if two tags have identical center co-ordinates, they possess the strongest connection possible and are in all likelihood to be synonyms or variations of the same word.

- **λ**
  The separating distance between two tags is proposed to be a function of the concentration of the tag cloud within which they reside. As such, these quantities are required to form part of a function which is used to compute probabilistic relations.

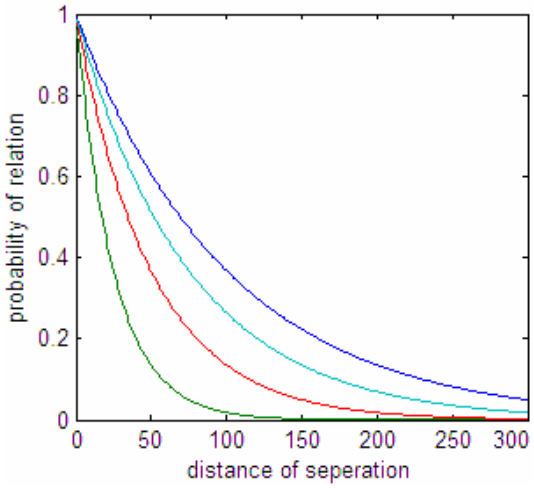
  

  Figure 24: Exponential decay functions with differing values of λ (green = 25, red = 50, cyan = 75 and blue = 100)

  The concentration of a tag cloud can be envisioned as how close the tags within it are to each other. If the distances between tags are small, then the concentration of tags within the cloud can be considered to be high. Similarly, if these distances are large, the cloud can be described as being scattered and dispersed. These observations require that a measure of concentration be created so that it may form part of the decay function.

  It was decided to take the average distance between the most frequent tag and every other tag within the cloud as a means to quantify concentration. The distances between the most frequent tag and every other tag were computed, summed, and then divided by the number of tags within the cloud minus 1 (because the most frequent tag is not compared to itself). This yields a value known as the *average separation*, and is particular to each tag cloud.

  λ is then assigned the inverse of the *average separation*, producing a graph similar to those presented above.

Explanation of the Graphs

The function plots on the graph presented above show the modified exponential decay function for different values of λ. An increase in the value pushes the graph to the left, while a decrease moves it in the opposite direction to the right. Changes in the function curve are also associated with these changes in value. As such, small values of λ, corresponding to a large *average separation*, give rise to curves which are similar to the blue curve above. Large values of λ, and therefore small *average separations*, produce curves such as the green curve, and lie closer to the origin.

As such, it is evident that the modified exponential decay function is suitable in transforming Euclidian distances of separation into probabilistic relations. Considering the green curve (*with a low average separation of 25*), tags which are in close proximity to one another receive a high probability of relation. As soon as the distance between the tags increases, however, the associated probability decreases drastically. This is a consequence of low *average separation*, and ensures that distances larger than this average yield low chances of connection.

An analysis of the curve in blue (*with a high average separation of 100*) suggests a different scenario. Large increases in displacement produce changes in probabilities which are much less severe than similar changes in the green curve would produce. As such, it is clear that the *average separation* is an active influence in determining the probability of connection between a pair of tags.

Shortcomings of the function

Although the proposed exponential decay function has been shown to be an appropriate means of modeling probabilistic associations between tags, there nevertheless exist areas of extreme that distort its output.

The graph presented above depicts a scenario where the *average separation* of a tag cloud is at its maximum possible value. This value is approximated by the hypotenuse formed by the sides of the tag cloud, and bounds the maximum distance that tags can be separated by.
As the *average separation* approaches this maximum value, the function begins to produce relatively high probabilities for this extreme distance that should in actual fact be

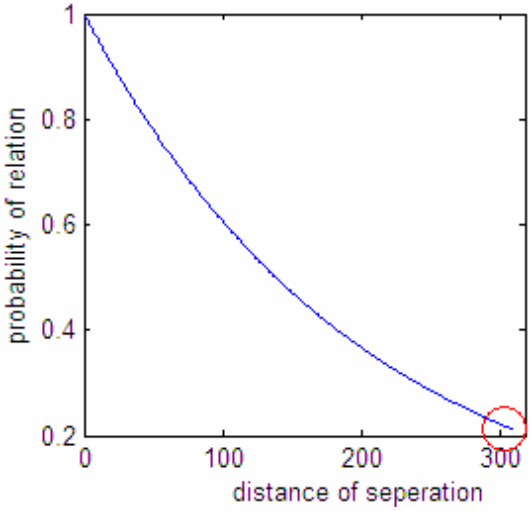


Figure 25: The modified exponential decay function, with the area of concern surrounded in red

close to zero. The graph shows that for a displacement of over 300 units, the associated probability is approximately 0.2, even though the tags can not be moved further apart to show their obvious lack of connection.

This event is highly unlikely, given the preconditions that are required to make it possible. In order for this occur, the most frequent tag must be positioned at a corner, while the majority of the remaining tags must be positioned at the opposite corner. This would yield the unusually high *average separation* that would give rise to such a situation.

8. INTERFACE EVALUATION [Refer to Appendix F]

While substantial effort has been committed to the evaluation of the data produced by the spatially-augmented tagging system (*see subsequent section*), consideration must be paid to the interface through which users are able to interact with it. As such, it is important to verify whether users of the system are able to effectively transform a non-specific cognitive understanding of term association to a corresponding spatial interpretation that is faithful to the former. If it is the case that users of the system are unable to do this, then the results of that they produce through their interaction with it are likely to be of a low quality. The results of this test are reproduced as Appendix G.

Ten test subjects were assembled and were asked to partake in a simple test. They were provided with a pre-existing tag cloud, where the tags were aligned to the left, and were asked to re-arrange the tags within it so that tags which they believed has a strong association are close together while those which have a weak connection are more dispersed.

The nature of the experiment makes it difficult to produce quantitative data relating to its results. As such, only qualitative statistics are reproduced here.

After the users had finished repositioning the tags, and were content with its appearance, they were posed with questions relating to their choices regarding the repositioning.

Figure 26: The tag cloud used for the Interface Evaluation

If two or more tags were in very close proximity to one another, they were asked whether they believed that a strong connection exists between them. All instances of this occurrence were probed, with all participants responding in a way that positively enforced their decisions.

If two or more tags were separated by a large distance, they were asked whether they believed that a very weak connection exists between them. Once again, participants responded, expressing their opinion that the distance that they put between the tags represented a weak connection between them.

Finally, the participants were quizzed regarding the ordering of tags within the cloud. For instance, if a tag or group of tags was position between another group, they were asked whether the middle group was connected to the tag or tags that surround it. They were then asked whether there was a weak connection between the groups that surrounded the center tag or tags. All participants responded confirming this, showing that they were able to re-order three dissimilar groups of tags so that their connection can be accurately expressed.

9. DATA EVALUATION [Refer to Appendix D]

The proposed extension of tagging practices, as is described within this report, is evaluated using several similar methods. The impressions of both expert and non-expert users are collected and compiled to provide a varied sense of analysis, while hands-on tests of the interface offer a more practical insight into the opinions of those that interact with the system interface.

9.1 Comparable Systems

The task of evaluating the tagging system and the data associated with its use proved to be exceedingly difficult. Although systems evaluation is a challenging and complicated task at the best of times, the process is made even more problematic when there exists no comparable implementation. While the author was initially certain that a similar extension of tagging as described within this report had been previously considered and implemented, many months of searching suggested that this was not the case. Consultation with academics from the department further supported this speculation.

The absence of a comparable system increased the difficulties faced when attempting to formulate a suitable evaluation of the spatially-enhanced tagging system. As such, a large part of the system assessment is based upon user feedback that relates both directly and indirectly to the software based tagging tool.

9.2 Regular User Evaluation [Refer to Appendix C]

9.2.1 Description of Evaluation

Regular users are defined as individuals that do not possess a knowledgeable understanding of the system in question and the technologies that are associated with it. More simply put, regular users can be envisioned as those individuals that casually make use of the Facebook application that communicates the ideas behind the extended tagging metaphor.

In order to assess the significance of the data created through the applications use, a questionnaire was compiled that related to three select tag clouds from the system [*See Appendix C*]. From these three tag clouds, the most frequent word from each was paired with every other word from that particular cloud. For example: if the most frequent tag from a hypothetical tag cloud was *cow* and the other tags from that cloud were *animal, milk* and *farm*, then (*cow, animal*), (*cow, milk*), and (*cow, farm*) would be the pairs that are created.

Three tag clouds were considered as it seemed likely that many people would be willing to complete a form spanning twelve pages. As such, the three chosen were selected based on their organization and structure, ensuring that they offered a representative illustration

of possible tag arrangements. These distinguishing factors are presented below and are referenced by their most frequent tag:

- *tea* cloud [Article 3]
  There is a definite central trend that is noticeable with this particular cloud. Although there is not a high degree of clustering, the majority of the tags fall within a region that appears to be at the center of the cloud.

- *astronaut* cloud [Article 8]
  Instead of being centrally located, the tags within this cloud are more dispersed, with tags appearing near all four corners of the cloud.

- *gas* cloud [Article 9]
  This cloud is unique in that there is a definite clustering of tags towards its center, while several tags appear to two adjacent corners.

Three lists of word pairs were created, as described above, with each pair being accompanied by a scale against which the participant could rate the strength of the word's connection.

Fifty test subjects completed the three forms (*one for each cloud*), and the results compiled into tables and graph that are shown in Appendix F:

9.2.2 Discussion of Results

Accuracy of Tagging and PLSA

The results of the test show a strikingly evident correlation between the results produced by the tagging application and those of the questionnaire. This trend is visually apparent when one considers the graphical representations of the data. However, there is a weaker relationship between the questionnaire data and the output of the PLSA algorithm, as is discussed below.

- astronauts
  The tagging and questionnaire bars for every word pair are extremely close together, and differ only marginally in many cases. The average difference between the probabilities produced by these two is only 0.12, suggesting a very strong correlation between them. A weaker correlation, however, seems to exist between the output of the questionnaire data and that of the PLSA algorithm. An average difference of 0.36 confirms this. This quantity is three times as great as the difference measured between the probabilities calculated through the questionnaire and tagging.

- tea
  The results of this comparison showed a similar trend to that which was observed in the *astronauts* test. There is nevertheless a slightly weaker correlation between

the results of the tagging and the questionnaire, while the results of the comparison between the PLSA algorithm and the questionnaire show a notable increase in correlation. Although this is the case, tagging again seems to provide a better mapping to the questionnaire data, with an average difference in probabilities of 0.19. While the PLSA algorithm fared better in this test, it still only was able to provide a difference in probability of 0.26.

- gas
  A similar overall trend seemed to occur in this comparison as well. Once again, tagging managed to produce a strong correlation between its output and that of the questionnaire, yielding a difference in probability of 0.15. Similarly, the PLSA algorithm produced results which indicate a weaker correlation to the questionnaire data, giving a high average difference in probability amounting to 0.39.

The results shown in Appendix F, along with the abovementioned statements, provide unquestionable proof that tagging yielded a stronger correlation between the probabilities that it generates and those that were entered by the test population. Tagging consistently produced results which were extremely close to those suggested by the test participants. Additionally, these results were always better than those produced by the PLSA algorithm as every test generated a difference in probability that was far smaller than what the automatic algorithm produced.

Another consideration is the number of times that each method produced the closest result to those of the questionnaire. Although this is related to the average difference in probability, it aids in highlighting the difference of these values.

Out of the thirty two word-pairs that were scrutinized by the test subjects, tagging proved to be closes to the results of the test on twenty six occasions (81%). As such, the PLSA algorithm was closest to the mark on the remaining six (19%). Further comparison between the tagging and PLSA algorithms are provided at the end of this section.

Affirming the Distance-to-Probability Transformation Function

The results presented above provide a means of support for the decision to model the association between distance and probability according to a modified exponential decay function. Although a function may exist that better describes this relationship, the results of the test aides in demonstrating that the chosen function provides results which are close to the perceived associations of the test subjects.

Affirming the Usability of the Interface

The results of the abovementioned test serves as further means to support the findings of the interface evaluation (*Section 6*). The strong correlation between the results of the test and the data produced by tagging show that users are able to express their opinions regarding connection in a spatial way. As such, the interface appears to afford the users

with a means to accurately convey their estimations of tag association to approximate distances which are analogue to this.

An Interesting Observation

An interesting trend was observed during the evaluation. After being questioned about the positioning of certain tags in relation to others, they justified their choice of location, and used *real-world* reasoning to support this. For instance, after asking a participant why he placed the terms *car* and *grass* in close proximity to one another, he stated that he understood that cars are found on top of grass, and no visa versa. What is interesting is that the *car* tag was placed above the *grass* tag to symbolize this relationship. Similarly, several users placed the tag *pen* above the *paper* tag, as they believe that pens are generally found on top of paper.

9.3 Expert User Evaluation [Refer to Appendix E]

To complement the results obtained from the evaluation of the regular users, expert users were asked to complete a similar questionnaire. Five MSc students from the department were asked to read all of the eleven articles from which the tag clouds were created. They were then instructed to select three words from the article that they thought provided the best summation of its contents (This question related to the output of the PLSA algorithm). In addition to this, they were asked to answer four questions relating to the tag clouds themselves.

9.3.1 Questions relating to the Tag Clouds

Questions were addressed to the expert users that attempted to probe their views and opinions regarding the arrangement of the tags within each tag cloud. These questions are presented below, along with statistics regarding the feedback returned.

i. Do you think that all of the tags that have been entered into the cloud provide a good summation of the article as a whole?

ii. Do you think that the tags which are in very close proximity to each other have a very strong connection?

iii. Do you think that the tag clouds are too cluttered, containing many infrequently used tags, as opposed to fewer tags of greater frequency? (A tag's size is related to the number of times that it has been added – the bigger its size; the more frequent its use.)

iv. Do you think that a linear relationship exists between the user's perceived connection between two tags and the distance that they place between them? (i.e. : if they are indifferent about the connection between the tags, would they place them half the distance of what they would if there was absolutely no connection between the two words at all?)

Question Number	Mean	Maximum	Minimum	Mode	Median
i	6.8	8	5	6	6
ii	6.4	9	4		6
iii	7.8	10	7	7	7
iv	3.8	5	2	5	4

Figure 27: A table showing statistics relating to the answers of the second part of the expert-user questionnaire

The first question asks whether all of the tags that lie with each tag cloud give a good summation of the related article. The response suggests that the test subjects believe that the tags provide a decent overview or the article, although it appears that *good* is too strong a term to describe the tags' ability to summarize the articles.

A similar view is apparent in the answers for the second question. It appears that the test subjects do not believe that a very strong connection exists between tags that are in close proximity to one another. However, the mean of 6.4 suggests that there does exist some positive correlation between a decrease in distance and an increase in connection.

The results of the third question strongly suggest that the tag clouds are compressed and dense, suggesting values of *average separation* that are below optimal. This could either be indicative of the former, or that users are choosing to enter their own tags than to rather re-enter existing ones.

The final question provides some support for the opinion of the author, reiterating the notion that a linear relationship does not exist between the distance that separates tags and their associated probability of relation. A maximum response of five suggests that even the most skeptical test subject was undecided, while the minimum response of two indicates strong support for the theory.

9.4 Tagging and PLSA Compared

The data gathered from the tagging application provides a means to evaluate the performance and connection between itself and the PLSA algorithm. Several aspects common to each can be compared in order to ascertain whether any relationship exists between their outputs.

9.4.1 Comparison of Generated Probabilities

All tags from each tag cloud were compared to all other words from the same cloud and their probabilities of association calculated. The same was done for the PLSA algorithm.

How PLSA determines probability of connection between terms

The PLSA algorithm generates a value for each term within a particular text, showing the likely hood that that particular word is the subject of it. For instance: if term X received a

probability of 0.54, then there exists a 54% chance that it is the subject of the article from which it forms part of.

In order to compare pairs of terms, the absolute value of the difference between each terms associated probability is calculated. This provides a means of associated probabilistic connection between these two terms. The formula used is presented below:

*probability of connection = 1 – (PLSA difference * 10)*

Figure 28: The formula used to convert PLSA output to a corresponding probabilistic relationship

Each value calculated using the above formula is graphed alongside the corresponding tagging probability, as shown below.

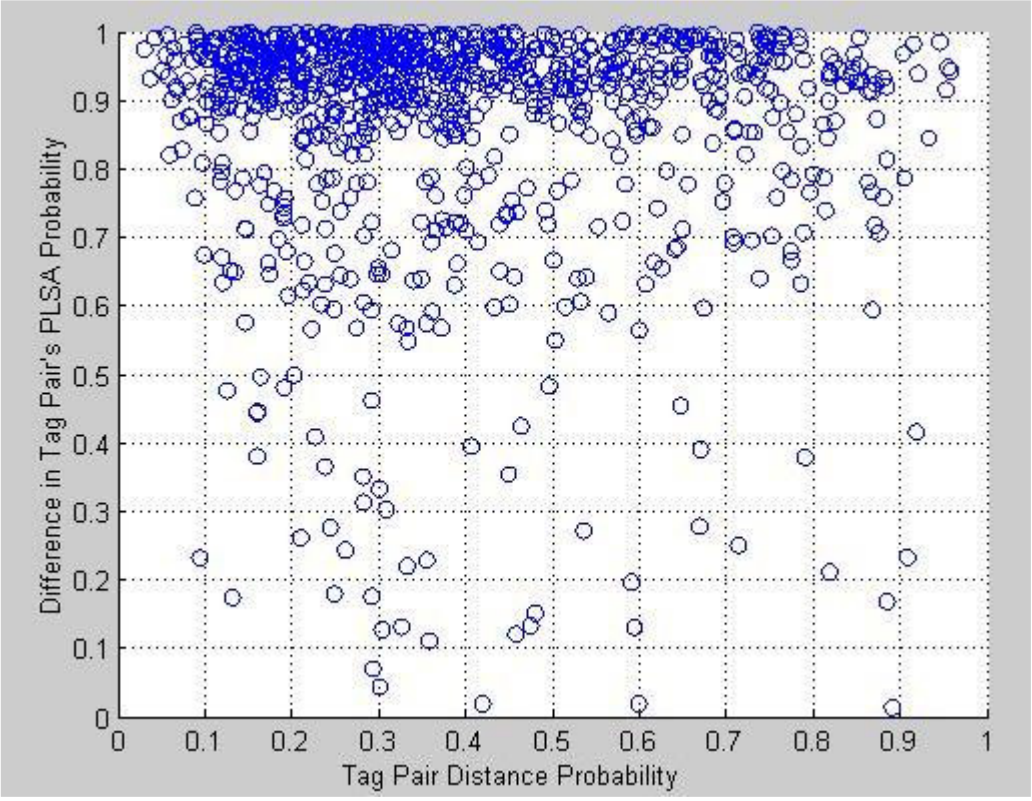


Figure 29: A graph showing the relationship between the generation probabilities of both PLSA and the tagging system

The graph shows no noticeable correlation between the probabilities generated by the PLSA algorithm and the tagging system. The majority of the outputs of the PLSA algorithm lie between the values of 0.8 and 1, allowing for little chance of correlation between these values and the varied values produced by the tagging system.

As such, no correlation can be made between the outputs of the PLSA algorithm and the tagging system.

9.4.2 Comparison of Term Frequency

A second comparison was performed between the outputs of the two algorithms. In this case, the tag frequency from the tag data was compared to the terms PLSA probability value. This assessment was made because these outputs attempt to approximate the same measure – the probability that a particular term provides the best synopsis of a particular text. A graph showing these comparisons is presented below.

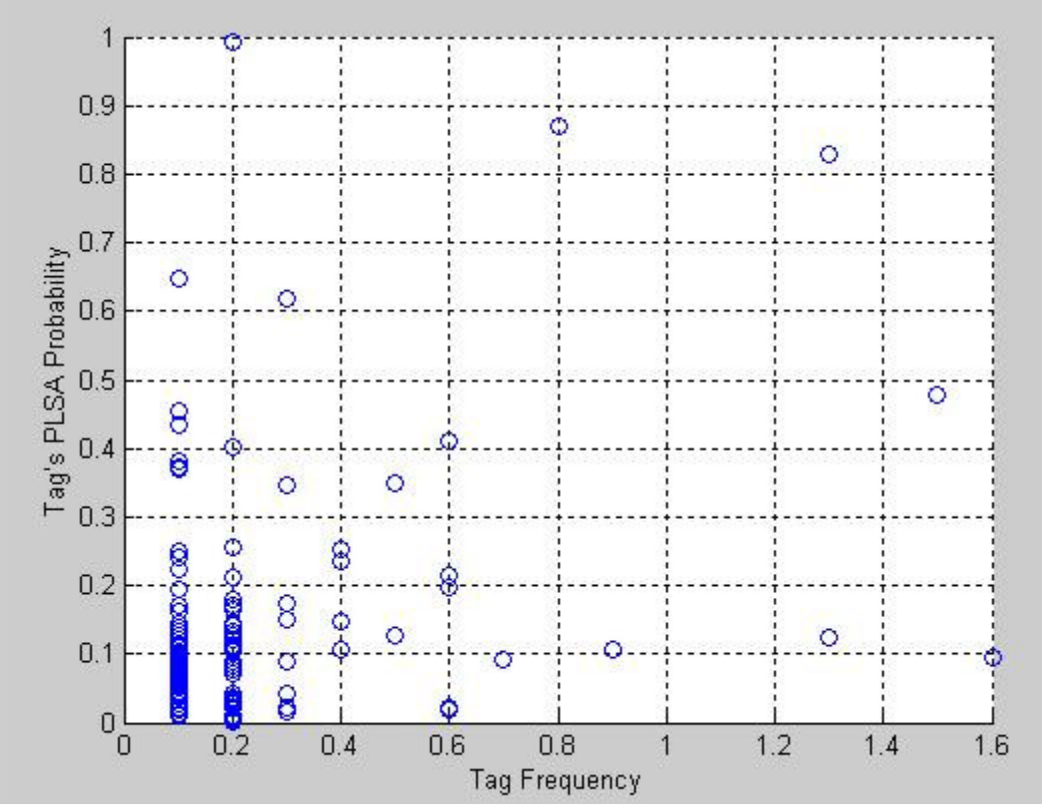


Figure 30: A graph showing the relationship between PLSA's output and tag frequency

Unfortunately, no distinct correlation can be found between the two measures. This is in all likelihood the low frequency of most tags. The vast majority of tags have a frequency of one or two, with only a few having ones greater than this. What is evident is that for extremely low probabilities, the outputs of the PLSA algorithm and the tagging system are partially aligned. This trend diminishes somewhat as the frequency increases.

9.5 Noticeable Hierarchical Structures

While the analysis presented above shows several practical and constructive aspects of the tagging system, there nevertheless exist certain tasks that it is unable to perform. It was initially postulated that spatially augmented tagging will not only provide a probabilistic association between words, but will also offer insight into the hierarchical nature of the tags that form part of each tag cloud.

The initial belief held by the author was that the most central term in the tag cloud would be the most general term, and therefore would have every other term within the cloud as an instance or subset of itself.

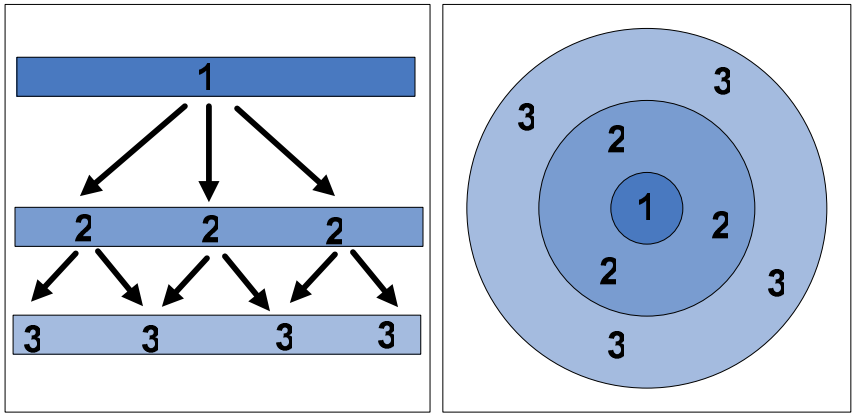


Figure 31: Images depicting the proposed hierarchical arrangements of tags within a tag cloud

The images above illustrate this theory. The circular image on the right represents a tag cloud with the most general term, **1**, in the middle. Immediately surround this term are other terms which are instances or subsets of the center tag, represented by **2**. Similarly, tags appearing on the outskirts of the tag cloud, represented by **3**, are subsets of both **2** and **1**.

While it was hoped that hierarchical data could be extracted from tag clouds, the structures presented above are not apparent in the tag clouds that have been generated through user interaction with the tagging system. Although through no fault of their own, users did not seem to position tags in such a way, ensuring that no definite hierarchical information could be retrieved.

10. XML OUTPUT

A major aspiration of the project was to produce a coherent, hierarchical knowledge structure such as a taxonomy or ontology. As stated in the preceding section, the addition of spatial importance to the tagging process did not generate data that allowed for these structures to be created.

While this is unfortunate, important probabilistic data was gathered, describing the relative strengths of connection between each tag and every other tag within its tag cloud. This information is useful as it shows a collective perception regarding the connections between particular terms, allowing for searching to provide results which are more current. This data was compiled into an XML document so that it may be used be others. An XML schema is included for completeness.

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

    <xs:element name="word" type="xs:string" maxOccurs = “unbounded”/>
        <xs:complexType>
            <xs:sequence>
                <xs:element name = “related-word” type = “xs:string” maxOccurs = “1”>

                    <xs:complexType>
                        <xs:sequence>
                            <xs:element name = “value” type = “xs:decimal”
                                        maxOccurs = “1”/>
                            <xs:element name = “probability” type = “xs:decimal”
                                        maxOccurs = “1”/>
                        </xs:sequence>
                    </xs:complexType>
                </xs:element>
            </xs:sequence>
        </xs:complexType>
    </xs:element>

</xs:schema>
```

Figure 32:An XML Schema created to transform the system’s XML output

In addition to this, the full list of tag pairs (909 of them) and their associated probabilities have been published online.

11. CONCLUSION

Although no hierarchical structures were able to be fashioned out of the data produced by the extended tagging metaphor, accurate and presentable associations between the terms of a close vocabulary were generated as a result of the systems use. Tests that probed a sample population's perceived associations between word pairs showed a strong correlation to the results of the tagging system. These results support the notion that

When compared to the results of a test that probed the test subjects' perceived associations between select words, the output of the tagging system appeared to show a strong correlation. This observation reinforced the notion that the spatial distances that separate pairs of tags is able to be transformed into a probabilistic relationship between the two bodies.

Comparison between the outputs of the tagging system and PLSA algorithm showed an unexpected lack of correlation between the two approaches, with the graphs suggesting an unknown random distribution. No function could be created to describe the relationship between the outputs of these two methods. Although this was unfortunate, performance of each algorithm could be measured against each other. The tagging method showed to be consistently closer to the inputs of the sample population.

In conclusion, the author believes that worthwhile investigation has been made into the merits of the suggested tag metaphor, while at the same time highlighting weaknesses of the approach. A greater number of tagging events would have certainly yielded results which would have allowed for the tagging and PLSA algorithms to be compared more thoroughly. However, the data that was gathered showed to a degree of certainty that distances between words can be transformed to represent probabilistic relationships.

12. FUTURE WORK

Several areas of improvement that are evident are certain to improve the effectiveness and value of the current system. These are presented below along with motivation for their inclusion.

12.1 Cross-Cloud Tag Consideration

There are several instances that are evident within the tagging data where identical tags were common among different tag clouds. Each of these individual tags has unique associations with the other tags that they are accompanied by within the tag cloud. At present, the data associated with each tag instances is not shared between all other tags of the same type. As such, there exists a means to magnify the possible number of tag pairs by sharing this disparate data amongst all instances of each type of tag.

12.2 An Improved Tag Movement Function

The weight function that forms part of the present solution is able to effectively weight the movements of every tag, ensuring that each change in position is scaled so that it exhibits a proportionate influence. Possible improvements exist, however, that may improve the success of this function.

One such modification takes the tag frequencies into consideration, and takes into account a form of gravity when calculating movements. Tags which are high in frequency will exert a passive attraction to all other surrounding tags in a manner that is analogue to the force exerted by all physical bodies in the universe. This addition could possibly aide in the creation of structures that are hierarchical in nature, as tag movements are continually influence by the frequency of all surrounding tags.

12.3 An Improved Automatic Means of Classification

The results produced by the PLSA algorithm were disappointing in some sense, although this is likely to be through no fault of its own. The documents against which the algorithm was run were short in length, thereby acting as a limiting factor to the algorithms performance. The inclusion of automatic algorithms that produce results which are more closely aligned to those of the tagging system may provide a means to automatically assess the *correctness* of tag placement. Unfortunately, malevolent users often attempt to disorder and confuse the state of the tag cloud by either entering tags which have no relevance to the content which they are scribed to, or by positioning tags alongside other tags that they have no connection with. Automatic techniques may aide in detecting these malicious actions before they are committed.

13. REFERENCES

[1] Leuf, P. *"The Semantic Web – Crafting Infrastructures for Agency"*. Wiley. 2006

[2] Tim Berners-Lee, James Hendler, Ora Lassila

[3] Mathes, A. *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. Unpublished paper.
Available Online: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[4] Quintarelli, E. *Folksonomies: power to the people*. ISKO Italy-UniMIB, Milan. June 24, 2005
Available Online:
http://www-dimat.unipv.it/
biblio/isko/doc/folksonomies.htm

[5] Borst, W. *Construction of Engineering Ontologies*. Centre for Telematica and Information Technology. University of Tweenty, Enschede, The Netherlands. 1997.

[6] Bateman, S. Brooks, C. McCalla G. *Collaborative Tagging Approaches for Ontological Metadata in Adaptive E-Learning Systems*.

[7] Bielenberg, K. Zacher, M. *Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation.* Masters Thesis. 2005.

[8] Xu, Z. Fu, Y. Mao, J. Su, D. *Towards the Semantic Web: Collaborative Tag Suggestions*.

[9] Golder, A. Humberman, B. *The Structure of Collaborative Tagging Systems*. Journal of Information Science, 32(2). 198-208. 2006.

[10] Halpin, H. Robu, V. Shepard, H. *The Complex Dynamics of Collaborative Tagging*. The 16th International World Wide Web Conference. May 8-12, 2007.

[11] Kroski, E. *The Hive Mind: Folksonomies and User-Based Tagging*. December, 2005.
Available Online: http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/

[12] Kuo, B. Hentrick, T. Good, B. Wilkinson, M. *Tag Clouds for Summarizing Web Search Results.* WWW 2007. May, 2007.

[13] Kaser, O. Lemire, D. *Tag-Cloud Drawing: Algorithms for Cloud Visualization*.

[14] Rivadeneira, A.W. Gruen, D. Muller, M. Millen, D. *Getting Our Head in the Clouds: Toward Evaluation Studies of Tag Clouds*. CHI 2007 Proceedings. April 28 – 3 May. 2007.

[15] *http://en.wikipedia.org/wiki/Stigmergy*

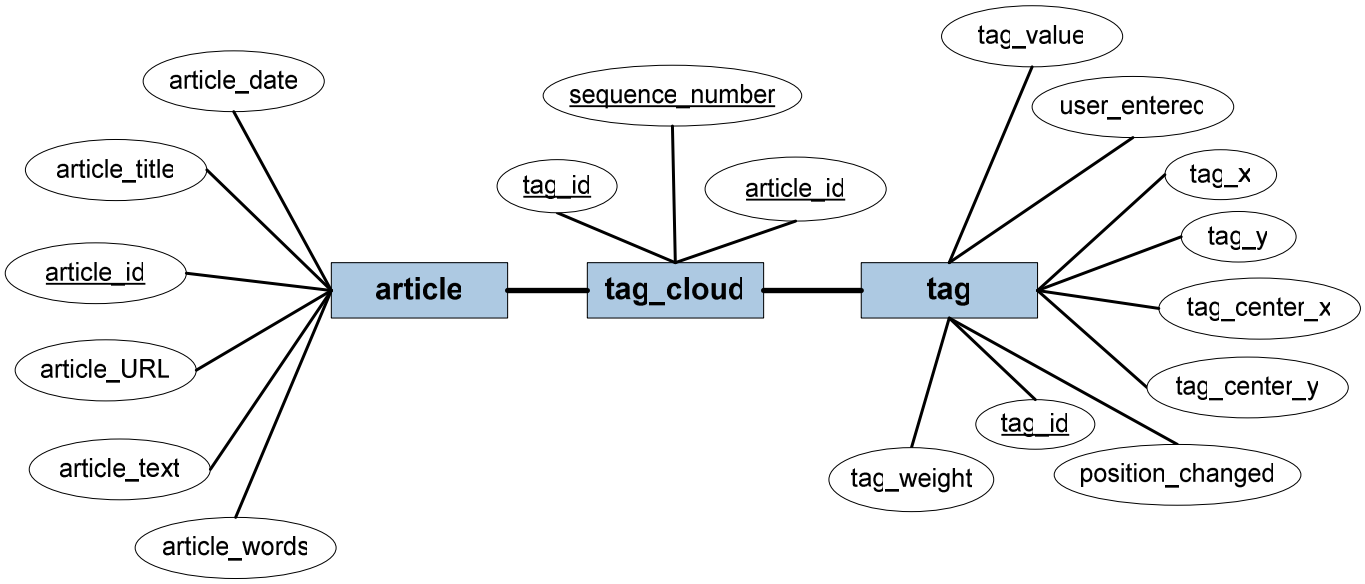
[16] Sebastiani, F., (2002), *"Machine Learning in Automated Text Categorization"*, ACM Computing Surveys, Vol. 34, No. 1, March 2002, pg. 1–47

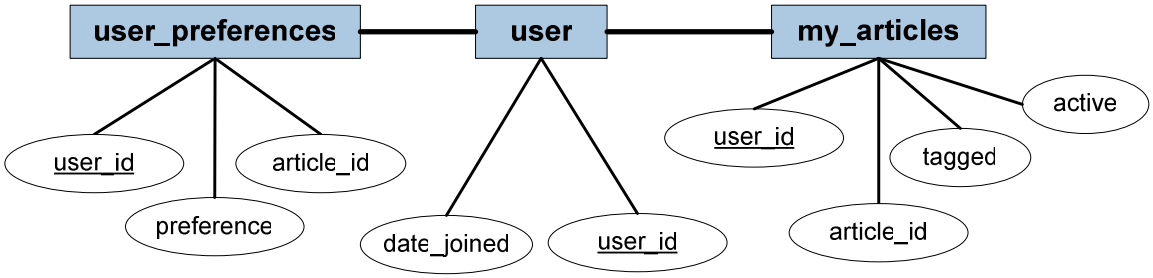
[17] *http://bitworking.org/news/Stigmergy*

[18] Stewart, J. *CalculusL Concepts and Contexts*. Brooks/Cole. 2001. p A9

14. APPENDICES

14.1 Appendix A – Entity Relationship Diagram





14.2 Appendix B - Final Tag Cloud States [Tag clouds have a width and height of 230 pixels]

Article 1

Value	Frequency	Position
alonso	13	93, 111
car	2	106, 135
emotions	1	145, 84
champion	3	64, 54
mclaren	2	54, 65
current	1	78, 58
hamilton	2	48, 51
wailing	2	111, 112
dirty	2	121, 113
squatting	1	142, 72
defeat	1	97, 144
rookie	2	43, 66
turd	2	150, 75
curling	1	20, 123

Article 2

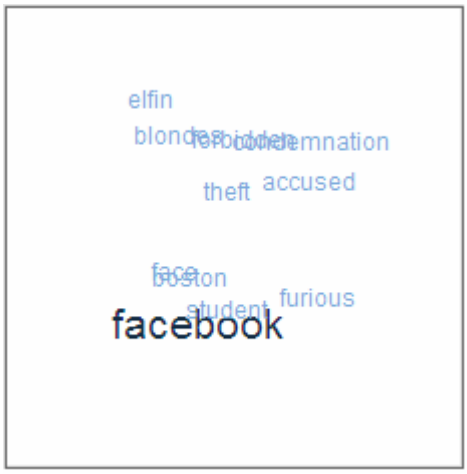
Value	Frequency	Position
unfunnyman	1	128, 44
stiller	1	83, 127
satan	1	122, 110
funniness	1	110, 110
comedy	1	121, 134
awful	1	98, 91
pact	1	102, 154
tricking	1	123, 155
cinematic	1	145, 154
sucker	1	84, 117
devil	1	66, 102
claw	1	72, 62
agony	1	91, 81
stinkbombs	1	138, 136
funny	1	54, 146
deliver	1	66, 102

Article 3

Value	Frequency	Position
tea	15	117, 104
diseases	1	136, 56
cancer	1	156, 46
cups	2	136, 111
sodomy	1	101, 80
britannia	1	84, 63
sarcasm	2	62, 97
eccentricity	2	85, 56
existence	1	117, 137
sunday	1	150, 151
nutrient	1	146, 66
lifestyle	2	58, 165
futility	1	10, 123

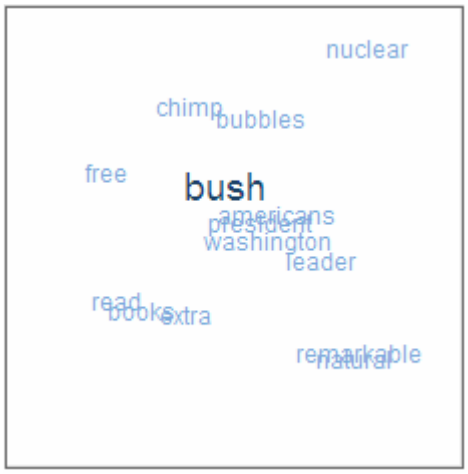
Article 4

Value	Frequency	Position
loser	6	113, 133
virgin	6	120, 117
spelling	2	92, 117
nyarlathotep	1	90, 143
champion	3	144, 68
tool	1	119, 130
prostitute	2	48, 21
condemnation	3	92, 161
booty	1	44, 14
cheerleaders	2	154, 51
sexually	2	57, 23
target	1	80, 101



Article 5

Value	Frequency	Position
facebook	16	78, 155
student	1	109, 153
theft	3	112, 92
accused	1	149, 89
blondes	1	84, 66
elfin	1	72, 48
forbidden	1	118, 68
face	1	84, 134
condemnation	3	148, 67
boston	1	91, 136
furious	1	154, 145
offenders	1	26, 123



Article 6

Value	Frequency	Position
bush	13	102, 85
president	2	126, 110
americans	2	132, 105
books	1	66, 154
washington	1	128, 118
chimp	4	90, 52
bubbles	1	126, 58
free	4	51, 83
extra	1	91, 156
read	2	54, 149
leader	1	155, 128
nuclear	2	179, 21
natural	1	174, 176
remarkable	2	174, 173

Article 7

Value	Frequency	Position
floods	7	77, 124
animal	1	17, 123
advertisement	3	93, 138
fernando	1	47, 29
poisonous	1	175, 161
almighty	2	120, 80
uk	3	72, 118
flood	2	68, 124
evan	1	100, 84
infected	1	190, 162
water	1	62, 131
original	1	60, 189
flying	1	112, 76
movie	1	104, 77

Article 8

Value	Frequency	Position
astronauts	9	92, 85
shuttle	3	91, 101
space	1	63, 95
doomed	1	187, 24
korean	1	168, 188
texas	2	42, 67
optimistic	2	53, 112
discovery	1	86, 85
safe	1	55, 182
foam	2	33, 188
trouble	1	193, 28

Article 9

Value	Frequency	Position
gas	8	76, 123
petroleum	1	89, 123
fuel	1	70, 121
gallon	1	82, 109
gasoline	3	102, 124
free	4	157, 78
suvs	2	31, 106
drivers	1	41, 85
price	1	161, 65
car	2	34, 101
pinball	1	148, 171
severe	1	190, 22
celebrated	1	170, 214

Article 10

Value	Frequency	Position
happiness	6	133, 105
australia	4	72, 66
penalty	1	173, 128
potato	6	128, 116
customs	1	70, 112
sydney	1	60, 61
parcel	2	73, 137
rugby	2	140, 100
ireland	1	84, 74
outlawed	1	161, 58
outlandish	1	27, 123

Article 11

Value	Frequency	Position
microsoft	6	85, 99
monkeys	3	165, 161
bill	1	79, 90
banana	1	159, 159
suspicious	1	84, 72
ms	1	76, 100
army	1	163, 28
employee	2	35, 190
peanuts	1	192, 172
related	1	19, 123

Appendix C – Regular User Questionnaire [Sample Size of 50]

1. **tea -> diseases**

1	2	3	4	5	6	7	8	9	10

weak association strong association

2. **tea -> cancer**

1	2	3	4	5	6	7	8	9	10

weak association strong association

3. **tea -> cups**

1	2	3	4	5	6	7	8	9	10

weak association strong association

4. **tea -> sodomy**

1	2	3	4	5	6	7	8	9	10

weak association strong association

5. **tea -> britannia**

1	2	3	4	5	6	7	8	9	10

weak association strong association

6. **tea -> sarcasm**

1	2	3	4	5	6	7	8	9	10

weak association strong association

7. **tea -> eccentricity**

1	2	3	4	5	6	7	8	9	10

weak association strong association

8. **tea -> existence**

1	2	3	4	5	6	7	8	9	10

weak association strong association

9. **tea -> sunday**

1	2	3	4	5	6	7	8	9	10

weak association strong association

10. **tea -> nutrient**

1	2	3	4	5	6	7	8	9	10

weak association strong association

11. **tea -> lifestyle**

1	2	3	4	5	6	7	8	9	10

weak association strong association

12. **tea -> futility**

1	2	3	4	5	6	7	8	9	10

weak association strong association

1. **astronauts -> shuttle**

1	2	3	4	5	6	7	8	9	10

weak association strong association

2. **astronauts -> space**

1	2	3	4	5	6	7	8	9	10

weak association strong association

3. **astronauts -> doomed**

1	2	3	4	5	6	7	8	9	10

weak association strong association

4. **astronauts -> Korean**

1	2	3	4	5	6	7	8	9	10

weak association strong association

5. **astronauts -> Texas**

1	2	3	4	5	6	7	8	9	10

weak association strong association

6. **astronauts -> optimistic**

1	2	3	4	5	6	7	8	9	10

weak association strong association

7. **astronauts -> discovery**

1	2	3	4	5	6	7	8	9	10

weak association strong association

8. **astronauts -> safe**

1	2	3	4	5	6	7	8	9	10

weak association strong association

9. **astronauts -> foam**

1	2	3	4	5	6	7	8	9	10

weak association strong association

10. **astronauts -> trouble**

1	2	3	4	5	6	7	8	9	10

weak association strong association

1. **gas -> petroleum**

1	2	3	4	5	6	7	8	9	10

weak association strong association

2. **gas -> fuel**

1	2	3	4	5	6	7	8	9	10

weak association strong association

3. **gas -> gallon**

1	2	3	4	5	6	7	8	9	10

weak association strong association

4. **gas -> gasoline**

1	2	3	4	5	6	7	8	9	10

weak association strong association

5. **gas -> free**

1	2	3	4	5	6	7	8	9	10

weak association strong association

6. **gas -> suvs**

1	2	3	4	5	6	7	8	9	10

weak association strong association

7. **gas -> drivers**

1	2	3	4	5	6	7	8	9	10

weak association strong association

8. **gas -> price**

1	2	3	4	5	6	7	8	9	10

weak association strong association

9. **gas -> car**

1	2	3	4	5	6	7	8	9	10

weak association strong association

10. **gas -> pinball**

1	2	3	4	5	6	7	8	9	10

weak association strong association

11. **gas -> severe**

1	2	3	4	5	6	7	8	9	10

weak association strong association

12. **gas -> celebrated**

1	2	3	4	5	6	7	8	9	10

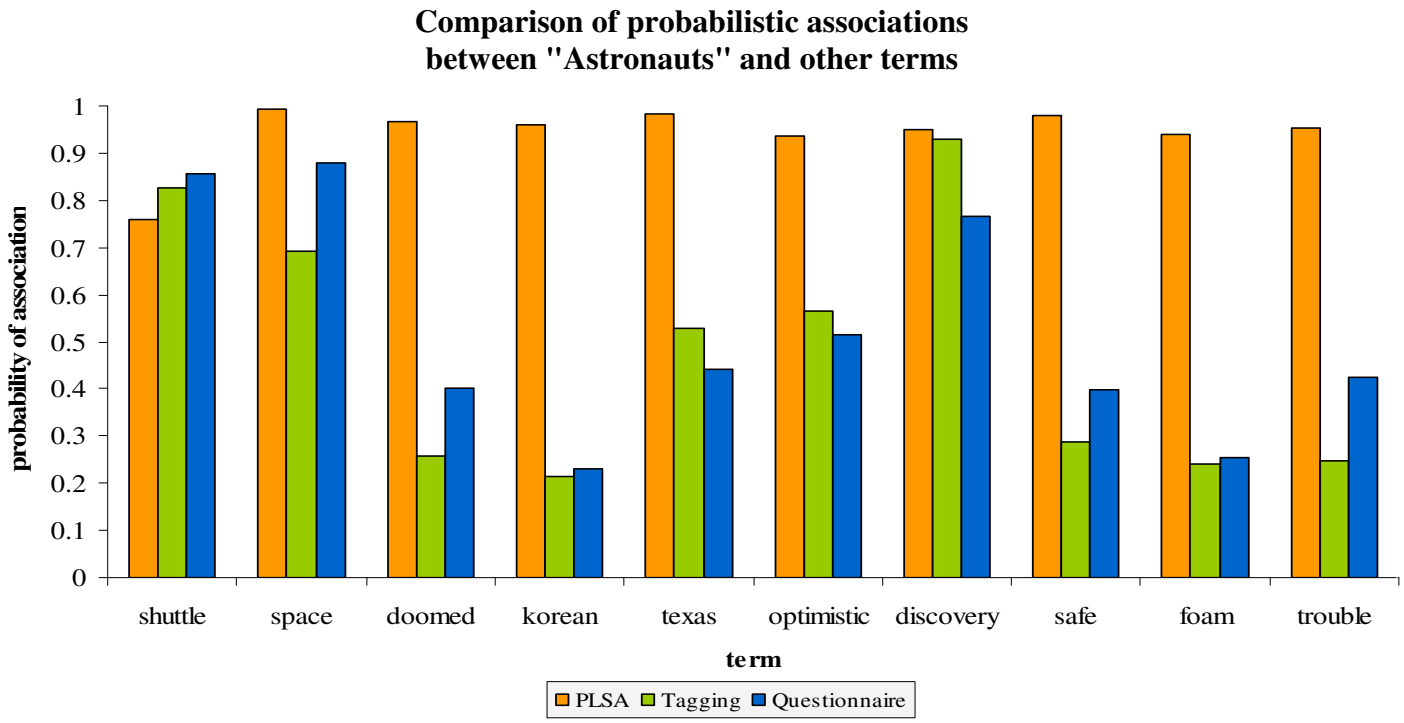
weak association strong association

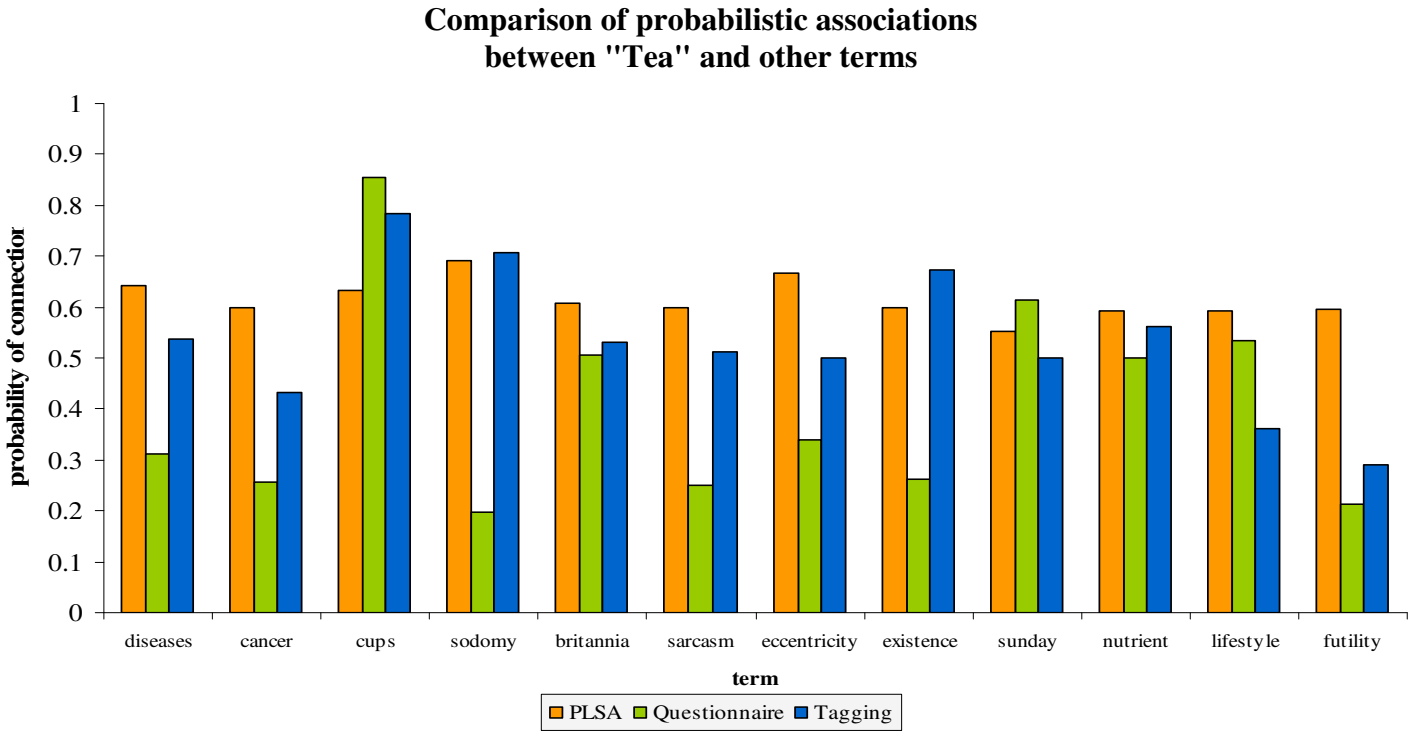
Appendix D – Questionnaire Results [Sample Size of 50]

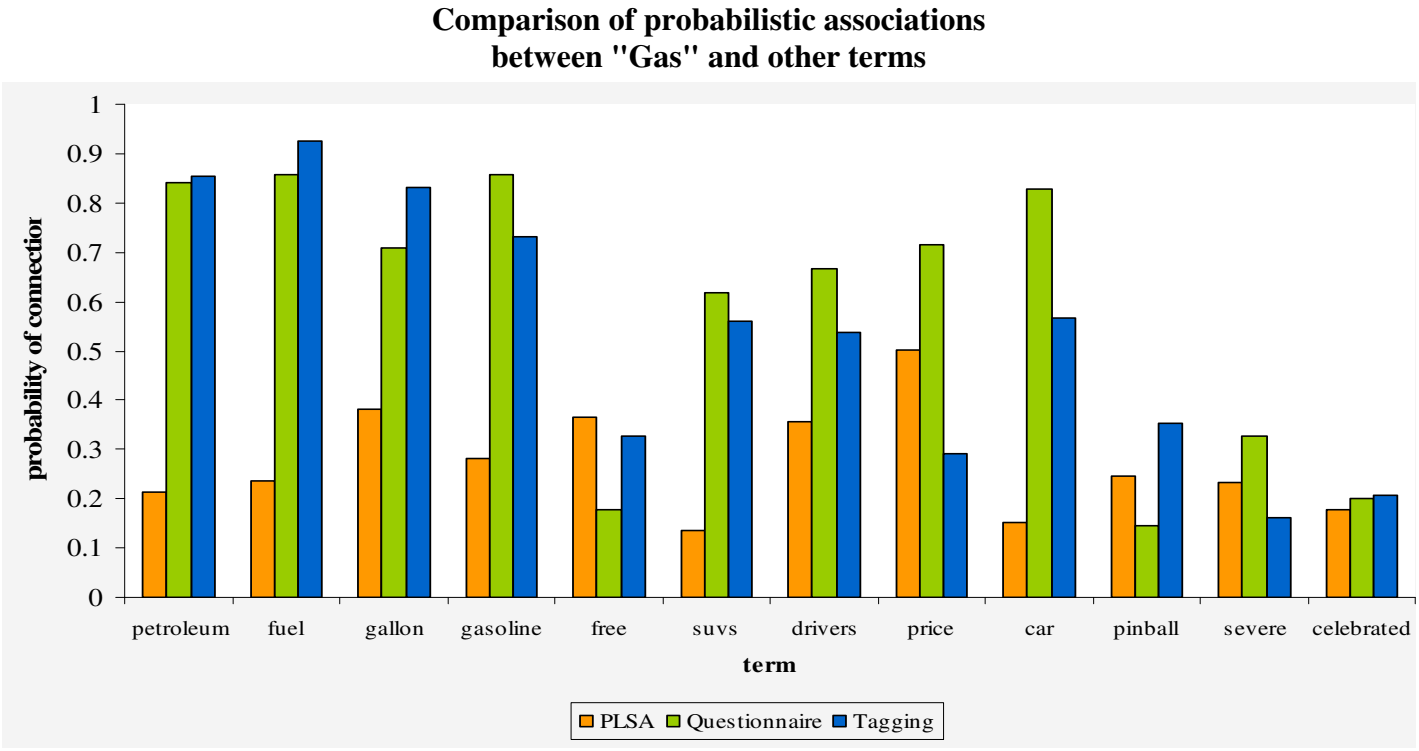
gas	Average	Maximum	Minimum	Std Dev	Median	Mode	Variance
petroleum	9.16	10	4	1.5826	10	10	2.5045
fuel	9.38	10	4	1.2271	10	10	1.5057
gallon	7.7	10	1	2.2246	8	10	4.9490
gasoline	9.36	10	5	1.3056	10	10	1.7045
free	1.82	6	1	1.4097	10	1	1.9873
suvs	6.06	10	1	3.2161	7.5	8	10.3433
drivers	6.6	10	1	2.5395	7	7	6.4490
price	7.68	10	1	2.4027	8	10	5.7731
car	8.86	10	1	1.8296	10	10	3.3473
pinball	1.44	10	1	1.3874	10	1	1.9249
severe	3.32	10	1	2.5026	3	1	6.2629
celebrated	1.94	10	1	1.8562	10	1	3.4453

tea	Average	Maximum	Minimum	Std Dev	Median	Mode	Variance
diseases	3.06	10	1	2.2804	2	1	5.2004
cancer	2.66	7	1	1.9755	2	1	3.9024
cups	9.26	10	1	1.4542	10	10	2.1147
sodomy	1.88	9	1	1.7337	1	1	3.0057
Britannia	5.2	10	1	3.2764	5.5	1	10.7347
sarcasm	2.62	8	1	2.2847	1.5	1	5.22
eccentricity	3.52	10	1	2.3925	3	1	5.7241
existence	2.38	10	1	2.1179	1	1	4.4853
Sunday	6.4	10	1	2.9137	7	8	8.4898
nutrient	5.22	10	1	2.3586	5	5	5.5629
lifestyle	5.48	10	1	2.2245	5	5	4.9486
futility	2.26	8	1	1.7706	1.5	1	3.1351

astronauts	Average	Maximum	Minimum	Std Dev	Median	Mode	Variance
shuttle	9.4	10	4	1.1606	10	10	1.3469
space	9.62	10	7	0.7796	10	10	0.6078
doomed	4.38	10	1	2.0790	4	4	4.322
korean	2.46	6	1	1.5012	2	2	2.2535
texas	4.6	10	1	2.7330	4.5	4.5	7.4694
optimistic	5.24	10	1	2.8468	5	5	8.1045
discovery	8.2	10	2	2.0702	9	9	4.2857
safe	4.22	10	1	2.3586	4	4	5.5629
foam	2.42	9	1	2.0413	1	1	4.1669
trouble	4.44	9	1	2.3226	5	5	5.3943


Comparison of probabilistic associations between "Astronauts" and other terms
probability of association
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
shuttle
space
doomed
korean
texas
optimistic
discovery
safe
foam
trouble
term
PLSA
Tagging
Questionnaire


Comparison of probabilistic associations between "Tea" and other terms
probability of connectior
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
diseases
cancer
cups
sodomy
britannia
sarcasm
eccentricity
existence
sunday
nutrient
lifestyle
futility
term
PLSA
Questionnaire
Tagging


Comparison of probabilistic associations between "Gas" and other terms
probability of connection
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
petroleum
fuel
gallon
gasoline
free
suvs
drivers
price
car
pinball
severe
celebrated
term
PLSA
Questionnaire
Tagging

astronauts	Average	Tagging Prob.	Difference	PLSA Prob.	Difference	Closest
shuttle	9.4	0.8248	0.1152	0.7591	0.1809	Tagging
space	9.62	0.6917	0.2703	0.9930	0.0309	PLSA
doomed	4.38	0.2575	0.1805	0.9672	0.5292	Tagging
korean	2.46	0.2148	0.0312	0.9614	0.7153	Tagging
texas	4.6	0.5280	0.0680	0.9837	0.5237	Tagging
optimistic	5.24	0.5655	0.0415	0.9370	0.4130	Tagging
discovery	8.2	0.9304	0.1104	0.9485	0.1285	Tagging
safe	4.22	0.2872	0.1348	0.9809	0.5589	Tagging
foam	2.42	0.2402	0.0018	0.9403	0.6983	Tagging
trouble	4.44	0.2482	0.2481	0.9524	0.5084	Tagging
			0.1202		**0.3573**	

tea	Average	Tagging Prob.	Difference	PLSA Prob.	Difference	Closest
diseases	3.06	0.5377	0.2317	0.6432	0.3372	Tagging
cancer	2.66	0.4317	0.1657	0.5989	0.3329	Tagging
cups	9.26	0.7840	0.1419	0.63301	0.2929	Tagging
sodomy	1.88	0.7070	0.5190791	0.6926	0.5046	PLSA
Britannia	5.2	0.5312	0.0112	0.6074	0.0874	Tagging
sarcasm	2.62	0.5136	0.2516	0.5997	0.3377	Tagging
eccentricity	3.52	0.4999	0.1479	0.6678	0.3158	Tagging
existence	2.38	0.6726	0.4346	0.5979	0.3599	PLSA
Sunday	6.4	0.5015	0.1384	0.5512	0.0887	PLSA
nutrient	5.22	0.5630	0.04103	0.5910	0.0690	Tagging
lifestyle	5.48	0.3606	0.1873	0.5925	0.0445	PLSA
futility	2.26	0.2908	0.0648	0.5951	0.3691	Tagging
			0.1946	**0.614**	**0.2616**	

gas	Average	Tagging Prob.	Difference	PLSA Prob.	Difference	Closest
petroleum	9.16	0.8553	0.0606	0.2141	0.7018	Tagging
fuel	9.38	0.9268	0.0111	0.2351	0.7028	Tagging
gallon	7.7	0.8327	0.0627	0.3806	0.3893	Tagging
gasoline	9.36	0.7314	0.2045	0.2803	0.6556	Tagging
free	1.82	0.3284	0.1464	0.3669	0.1849	Tagging
suvs	6.06	0.5609	0.0450	0.1356	0.4703	Tagging
drivers	6.6	0.5375	0.1224	0.3559	0.3040	Tagging
price	7.68	0.2903	0.4776	0.5008	0.2671	PLSA
car	8.86	0.5656	0.3203	0.1535	0.7324	Tagging
pinball	1.44	0.3535	0.2095	0.2457	0.1017	PLSA
severe	3.32	0.1603	0.1716	0.2340	0.0979	PLSA
celebrated	1.94	0.2075	0.0135	0.1766	0.0173	Tagging
			0.1538		**0.3855**	

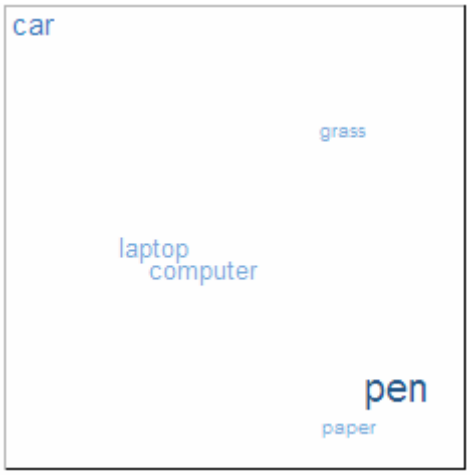
Appendix E – Results of the Expert User Questionnaire [Sample size of 5]

Article	1st Choices	Tag Frequency	2nd Choices	Tag Frequency	3rd Choices	Tag Frequency
1	champion	3	mclaren	2	hamilton	2
	more	0	protest	0	champion	3
	driving	0	emotions	1	Alonso	13
	world	0	hamilton	2	mclaren	2
			champion	3	protest	0
2	unfunny	1	comedy	1	audiences	0
	movies	1	deliver	0	movie	1
	devil	1	stiller	1	comedy	1
	hollywood	1	movie	1	schrick	0
	painfully	1	unfunny	1		
3	engrish	0	eccentricity	2	genes	0
	killer	0	diseases	1	diet	0
	diet	0	tea	15	lifestyle	2
			engrish	0	drink	0
			chemical	0	vitamin	0
4	achievement	0	condemnation	3	despaired	0
	champion	3	word	0	fantasy	0
	kid	0	champion	3	loser	6
	smart	0	kid	0	ban	0
			loser	6		
5	property	0	theft	3	risk	0
	theft	3	interweb	0	details	0
	facebook	16	condemnation	3	face	1
	interweb	0			victim	0
					minority	0
6	freedom	0	read	2	write	0
	president	2	president	2	laws	0
	read	2	linguistics	0	learn	0
	learning	0	natural	1	learning	0
					tricked	0
7	boat	0	water	1	cinema	0
	movie	1	washout	0	promotion	0
	flood	2	advert	0	cinema	0
	england	0	movie	1	advertising	3
	doomed	0	safety	0		
8	explosion	0	guaranteed	0	burned	0
	shuttle	3	mission	0	explosion	0
	space	1	doomed	0	fire	0
	doomed	0	safety	1	death	0
					advertising	0
9	consumers	0	complaints	0	ceased	0
	price	1	petroleum	1	advertising	0

	gasoline	3	free	4	waste	0
	usa	0	gas	8	price	1
	gouging	0	enforce	0	backlash	0
10	facial	0	concealment	0	importing	0
	toy	0	genitals	0	parcels	2
	happiness	6	happiness	6	outlawed	1
			toy	0	imprisonment	0
			arrests	0		
11	typewriters	0	monopoly	0	computers	0
	monkeys	3	keyboard	0	windows	0
	windows	0	microsoft	6	bananas	1
	banana	1	computers	0	monkey	3
			surprised	0		

Words with a frequency of two appear in blue, and those with in green.
Frequency is defined as the number of that a particular term was chose as either a first, second, or third choice for a particular article.

Appendix F – Results from the Tagging Interface Test [Sample size of 10]


car
grass
laptop
computer
pen
paper