Exploring Crowdsourcing to Personalize Web Experiences

Thesis submitted in partial fulfillment
of the requirements for the degree of

MS by research
in
Computer Science & Engineering

by

Deepti Aggarwal
201007001
deepti.aggarwal@research.iiit.ac.in

Software Engineering Research Center
International Institute of Information Technology
Hyderabad - 500 032, INDIA
November 2013

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled *"Exploring Crowdsourcing to Personalize Web Experiences"* by *Deepti Aggarwal*, has been carried out under our supervision and is not submitted elsewhere for a degree.

_____
Date

_____
Advisers: **Prof. Venkatesh Choppella**

_____
Prof. Vasudeva Varma

To my loving parents

RAKESH AGARWAL and NEERA AGARWAL

and to a special person

ROHIT ASHOK KHOT

Acknowledgments

First and foremost, a sincere thanks to my adviser *Prof. Venkatesh Choppella*. This thesis is the outcome of his excellent guidance and his constant backing. I am really thankful to him to give me complete freedom to explore my interests throughout my degree. His cheerful nature and energy have always motivated me to do something new. And on a lighter note, Sir the coffee sessions and the meals shared with you were really helpful, sometimes to build up my enthusiasm and some other times to shed my work load...

I also thank *Prof. Vasudeva Varma* for his invaluable guidance and support to accomplish my work. A special thanks to *Prof. Anind K. Dey* under whom I have taken a course on Human Computer Interaction. The course had indeed changed my outlook to perceive technology as a helping hand for human beings.

Special thanks to *Dr. T. B. Dinesh* to provide me an opportunity of working on Alipi, which has been a nice and learning experience for me. I am also grateful to *Prof. K. Viswanath* and *Mrs. Kavita Vemuri* for their guidance.

Thanks to my lab mates *Himanshu*, *Manjula*, *Sai* and *Swathy* who were always there to lighten my mood, to listen to my ideas, to review my work and to provide me valuable suggestions. Many many thanks for your guidance and support. I also thank Virtual labs team *Bodhi*, *Swetha*, *Mahesh*, *Vamsi* and *Medhamsh* for helping me whenever required.

I appreciate the contribution of every participant who has been a part of the experiments conducted related to my work. Their cooperation and feedbacks have made this research successful.

Thanks to all my colleagues *Swagatika, Ruchi, Mahathi and SRK* and my best pals *Sourabh, Sunny, Monika, Nainpriya* for holding me in my bad times. I am thankful to my family for their understanding and faith on me. This work would not have been possible without their unwavering support and love. And lastly, I am left with no words to thank you *Rohit* for keeping the flame enlightened in me. Thanks for being my inspiration!

Abstract

With the rapid advancements of communication and computational technology, the World Wide Web has witnessed a rapid growth in the user-generated content with more and more users actively creating, publishing and sharing content over the web. As a result, the web is now overloaded with information on varied topics contributed by diverse set of contributors. This phenomenon has given rise to the "big data" wherein lies a key problem of intelligently extracting the most relevant and accurate information specific to a user. With this, it is essential to provide more personalized web experiences to the user, where every user query is catered and satisfied according to her preferences. However, doing so would require extraction and understanding of the context and semantics of the content, which currently is not readily available. Moreover, automated systems show limited capabilities in performing the same task.

This thesis is an attempt to utilize collective human intelligence to support extraction and understanding of the content over the web, which will in turn help to create personalized web experiences. In particular, we propose crowdsourcing based systems for the following tasks: 1) extracting user preferences, 2) extracting named entities, and 3) renarration of the web documents. First, we propose a friendsourcing based approach called as *Crowd Consensus* where we extract user preferences from the collected opinions from her friends and tested it with an online game called as *Power of Friends*. The current method of eliciting information is to pose direct questions to friends and expect a truthful response in return. Power of Friends, on the other hand, involves a novel way of identifying the unanimous opinion of all the friends about a question related to an individual. Next, we describe a system called as *uPick*, which extracts named entities and their relations from a given text and crowdsource these extracted named entities for validation. The existing systems built around the task of identifying associated relations among named entities within a text document lack human precision and they also struggle to handle erroneous documents. uPick helps to improve the accuracy of the generated relations by gathering judgments from the interested users and validate the relations based on the majority responses. Finally, we worked on a renarration framework to the web called as *Alipi* to make the multi-lingual documents accessible to the users. This framework supports alternative descriptions for a web page or parts of it via rewriting for a given target audience by volunteers. We developed a browser plugin to enable users to re-narrate any page and to render the requested page dynamically based on the user preferences.

Our developed prototypes along with the studies show that leveraging human energy and skills have potential to provide solution to the problems that machines cannot accomplish solely. We hope our work would inspire system designers to consider crowdsourcing based systems for creating personalizing web

experiences and to think beyond system efficiency and accuracy by focusing on the task experience and invested efforts by the users.

Keywords: Crowdsourcing, Human computational games, Friendsourcing, Community knowledge, Objective facts, Web accessibility and Re-narration web.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The evolution of personalized web

Back in 1993, when World Wide Web services were starting to surface, people envisioned web as being one big information portal where there will be access to variety of different information on variety of different topics (Figure 1.1). The focus was on ownership and publishing of the content, written by someone, with an ability to browse and search content. The web was full of static pages and the users were predominantly content consumers. For example, with Web 1.0, a web user searching for good hotels in Shimla had access to all the static web-sites advertising or providing information about different available hotels in Shimla with their price and living conditions. The overall goal was to present information to the users as a catalogue and attract the customers. With the advancements in communication technologies around 2003, the web was started to be considered as a social platform, where the focus was on producing and sharing content. The Web 2.0 also introduced social media services like Facebook, Twitter that facilitated easier and richer communication among people. Here, instead of having dictated by someone, Web 2.0 was socially constructed and consumed. Therefore, with Web 2.0, the users were able to read about the experience of other customers in different hotels at Shimla and interact with them through blogs, and social networking websites to find the best suitable hotels. The web pages became dynamic with improved user interaction. People as a result took a leap from being content consumers to content producers and have given rise to huge user generated content (UGC)[17]. As a result, the web is now overloaded with information of different domains from different contributors.

Now that internet has already penetrated the households and everyone now enjoys an uninterrupted access to a variety of different web services, focus of the Web is slowly shifting towards making it personalized and portable. The onus is on building interactive prototypes to facilitate convenient, contextualized search and personalized web access. For example, giving users the opportunity and options to search and access content that match their preferences. However, to create a personalized web, the web should be capable of reading and understanding the user context and provide information to the user by considering his preferences for which every data should own its semantics and context of the

1990	2003	2010	2020
Web 1.0	**Web 2.0**	**Web 3.0**	
Web as an information portal	Web as a social platform	Web as a personalized portable web	
Focus on ownership	Focus on community	Focus on individual	
Static web pages	User generated content	Semantic and portable pages	
Meaning is dictated	Meaning is socially constructed	Meaning is socially constructed and contextually reinvented.	

Figure 1.1 Evolution of the web from Web 1.0 to Web 3.0

content for a user is defined by the data itself. Here, 'context' refers to why the content is relevant and to whom while 'semantics' refers to the meaning of data and how it is relevant to a given context. Once we have extracted both context and semantics for every available data, we envision that the web will be able to deliver the following applications.

1. Personalized browsing experience

Browsing experience of the users can be enhanced by considering his preferences and choices. If the user context is known, then the information and advertisements can be provided accordingly to the user. A website knowing the user's location and buying habits, will present offers and suggestions suitable to his interests. However, to make the website aware of user context, creating a user profile containing his preferences and choices is necessary, which is the main motivation behind building personalized web.

2. Data-on-demand

Once we have all the documents semantically connected to each other, every information request, complex or simple, can be satisfied in an efficient and accurate way. Unlike the existing search, which requires tedious browsing in order to get to the desired content, it will become much easier and faster to serve every request of the user. For example, assume that someone is interested to write an article on the glorious career of Sachin Tendulkar and wants to know more about his school and childhood days. In the present scenario, it would require parsing through and reading 'n' number of articles available on the web related to Sachin. However, if all those articles were semantically connected to each other, then it would have been much easier for the individual to fire a direct query such as "what is the name of the Sachins primary school?, When and where did Sachin play his first match?", and the search engine would have provided an exact answer to the same.

3. Multi-lingual web

Currently, the web documents are mainly dominated by English language documents (57%) with other languages as German, and Japanese forming 5% of web documents, and all other languages along with around 1600+ Indian languages fall in the remaining 10%. Given this web page distribution, it is evident that a non-English speaker will find it difficult to understand and consume information available in English. For example, a child from Karnataka is looking for fire safety related information such as how does the fire bus looks like, what steps should be taken if there is a fire accident in school, what are the first-aid measures, in his native language, Kannada. When he browsed the web, there are plenty of pages available on fire safety, however, their availability in Kannada is very less. This problem could have been solved if the content available in other languages is made accessible to the child by translating the relevant documents in Kannada, or by providing him web sources in different formats (such as videos or images) or sources (such as blogs or personal pages) but available in Kannada.

We envision that once the web documents are semantically linked with each other such that all the web sources in different languages and formats related to the same topic are connected, and the user context is known, creating the web as a personal tool is achievable.

1.1.1 Challenges to achieve personalized web

In order to satisfy and reach above-mentioned applications of personalized web, it is essential to extract and understand context and semantics behind every documents. However, we identify three key challenges in doing so, which we mention below (Figure 1.2).

1. Scattered data

Most of the data on the web is scattered and not explicitly available in form of documents. For example, social network like Facebook contains tremendous amount of information related to the user: his interests, places visited, languages known, and areas of expertise. However, this useful information is available only within his community and is shared by his community members. If this data is available, personalizing web experiences becomes an achievable goal. Therefore, the major challenge is of finding efficient ways to extract and validate this scattered data from the activities of a community.

2. Excess of data

With Web 2.0 and proliferation of UGC, the problem is not the sparsity of the content rather the excess of it. Current social platforms like blogs, wikis as well as social networking sites provide users with easy way to publish content on the web. As a result, the content is available on diverse topics and from variety of different sources. Identifying exact and best suitable content thus remains a major challenge, which becomes focus of the thesis.

3. Understanding data

The data is often distributed in variety of different sources and formats over the web. Since each source

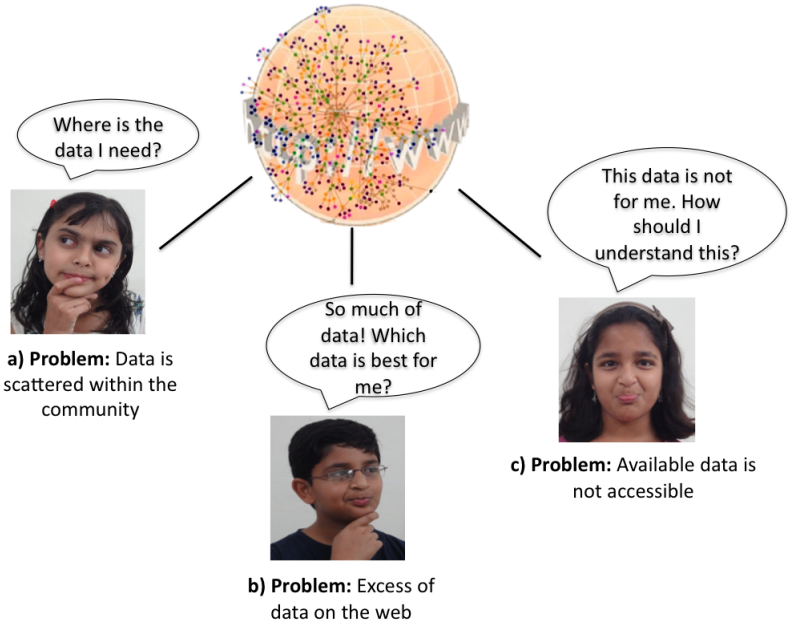

Figure 1.2 The figure shows three key challenges related to personalized web that became the focus of the thesis: a) Acquisition of the scattered and sparse community related data. b) Identifying the best suitable information from millions of online documents. c) Understanding the data available in different langauges.

follows its own language and structure, it becomes difficult to conceive its meaning, especially, if the languages and the structure are unknown. Therefore, the challenge is to create an understanding of the available data.

4. Limitations of the automated techniques

Although automated techniques are faster and provide better performance, however, for certain tasks such as validation and filtration of the data, automated systems can achieve only limited success. The reason being their dependency on specific domains, and datasets, which make it difficult for an automated system to scale well when applied in different scenario. Therefore, there lies a need to explore an efficient approach so as to provide solution to the tasks where autoamted techniques fail to function.

I believe exploring these challenges are essential for designing personalized web experiences for the web users. I envision that the thesis will contribute the following:

1. An explanation of how crowdsourcing can help to solve above mentioned problems.

2. Three design prototypes that explores each of the following problem using crowdsourcing.

3. A conceptual understanding developed through analysis and user studies.

1.2 Our Approach: Bring humans in-the-loop of automated systems

In order to solve our research problems of personalizing browsing experience, providing data-on-demand, and creating a multi-lingual web, we initially looked at the available automated techniques. Undoubtedly, the automated systems are capable of doing longer and repeated computations with high efficiency. For example, machines show high performance to extract all the named entities and their relationships from a huge input corpus of documents. However, despite the advancements made in machine learning techniques and artificial intelligence algorithms, the level of accuracy that these algorithms can achieve in the extraction and translation still unmatched with the humans. It has been found that the automated techniques can achieve limited success in a specific setting such as in particular domain or dataset. Therefore, these algorithms suffer with problems of scalability and cost effectiveness along with accuracy. Hence, it becomes essential to include humans in the working loop of the automated systems. Although humans are far better in performing tasks which require cognitive skills, however, humans find such computational tasks as non-engaging and thus require some incentives to participate. The focus of our thesis is to explore, how to invite human participation to perform high computational tasks.

We have also applied the three central aspects of human computation systems, that is, What-Who-How, proposed by voh Ahn to perform a human computational task[68]. The three aspects provide a wholistic view to solve a problem by combining both human and machine intelligence: "What" aspect decides the operations (decomposition of task into sub-units and then aggregation of outputs to form final answer) and their sequence to be performed to accomplish the assigned task; "Who" aspect defines the assignment of each operation to either machines or human workers; "How" aspect pertains to design systems that could invite human participation and motivate them to perform best of their abilities. To solve the above mentioned research problems of our thesis, we also define three aspects of the problem to determine who(human or machine) will perform what phase of the system and how to accomplish that phase. We tried to identify which part of the system requires human intelligence to improve its accuracy and how it should be achieved. For example, the existing human computational games built around the task of extracting knowledge from the web sources show that humans are efficient in both extraction and validation tasks. However, it has been observed that humans perform better in validation task because the extraction task often tends to be task oriented and thus requires higher cognitive efforts[68]. Considering the overall performance of a system, we have accounted both automated techniques and human efforts and determined appropriate stages of the system for applying either of the approaches. For example, in our proposed approaches, computationally high task of extraction is performed by the machines and validation of the extracted data is performed by the humans.

1.3 Thesis Statement

Located within the HCI tradition, this thesis explores the features of crowdsourcing based system to support creation of personalized web experiences. In particular, this work involves investigation in the following three spaces: 1) crowdsourcing 2) automated techniques 3) personalized web.

The main research question explored in this thesis is: *"How to design crowdsourcing system to personalize web experiences for the users?"*

I will utilize research through design[110] and mixed method research[30] practices to answer the research question of the thesis. Considering the broadness of the topic, I have focused on the following three problems.

1. Extraction of user preferences

With the advent of social web, much of individuals information including his preferences is now available online, particularly, on social networks. These preferences are particularly useful in personalizing individuals browsing experience and to support targeted advertising. However, they are often scattered across various web pages and are hard to extract and validate. To solve this, we propose a friendsourcing based approach called as *Crowd Consensus* where we extract this useful information about an individual by asking his friends. Based on the proposed approach, we developed a human computational game called as *Power of Friends* to make the task fun and engaging.

2. Extraction of named entity relationships

In order to facilitate semantic connection between the documents, we need to consider language constructs of the document. We have considered English language for our exploration, where the atomic unit of text is called as Named Entity. These named entities are of different types such as noun, verb, adjectives, which when linked together, form a relationship. However, given the diversity and the excess of this information, automated techniques are not efficient in extracting these named entities from the available documents. To solve this, we propose a crowdsourcing based approach where we utilize the human intelligence for extracting named entities relations from the document. We developed a system called *uPick* that use POS taggers to extract named entities from a given text and crowdsource these extracted named entities for validation.

3. Re-narration of the web documents

In order to make the webpage, available in multiple languages, accessible to the user, we need to provide the web-pages to the users in their prefered languages. To achieve so, we have worked on a Framework called as *Alipi*. Alipi follows a re-narration approach to solve the accessibility issue with the webpages, where every web user is allowed to re-narrate any element of a web page (as a re-narrator) and any user (reader) is allowed to see the available re-narrations for the queried page in his prefered language. Re-narration here implies rewriting of an elements in the DOM architecture of the webpage with a new content such that these re-narrations are avilable as different versions of the same page targeted

for a specific community sharing common preferences. We tested the first web-based prototype and developed a browser based plugin to make the task of re-narrating web-pages easier for the users.

1.3.1 Research Objectives

The research objectives to achieve our research goal are described as below.

Objective 1: Gather understanding of the relationship between machine and crowdsourcing systems from the existing literature.

The thesis will enumerate existing systems, both automated and human computational, for extracting personal preferences from the community activities, extracting named entity relationships from the web documents and making the webpages accessible for a user with certain language preference. It also includes issues and opportunities associated with machine and human computation to achieve goals of the thesis.

Objective 2: Explore the design space of crowdsourcing systems in relation to creating personalized web.

This thesis will present three design prototypes built upon the identified design opportunities associated with personalized web. These design explorations will serve as research vehicles to develop insights into how to develop crowdsourcing based systems for solving the similar problems.

Objective 3: Validate the design space.

This objective involves an empirical evaluation of the experiences with the designed prototypes. Analysis of the user experience, gained from the studies, will help us to formulate an understanding of how crowdsourcing can impact the experience of user related to our goals.

1.4 Research Methodology

We will utilize research through design and mixed method research practices to answer the research question of the thesis, which we describe below.

Research through design

Research through design is a reflective practice where thinking occurs through prototyping to examine the process, invention, relevance and extensibility of the design[110]. This approach helps interaction designers to integrate models and theories with the technical opportunities demonstrated by engineers to make the right artifact to transform the world from its current state to a preferred state.

Mixed method research

Mixed method research focuses on collecting, analyzing, and mixing both quantitative and qualitative data in a single study or series of studies[30]. Mixed method research is advantageous for these reasons: 1) Variation in data collection leads to greater validity. 2) Mixed method research answers the question from a number of perspectives.

Quantitative research collects objective data with predetermined instruments that yield statistical validity while qualitative research involves collection of open ended subjective data with an intent of developing themes[30]. We will use questionnaire and likert scales responses for extracting quantitative data from the participants about the designed systems. We will also utilize interviews method to collect the subjective data about the experiences of participants with designed systems. Interviews are commonly used to understand the mediated interactions of a user with technology. Interviews provide insights on the user experience that cannot be measured through quantitative data. We will take notes during the interview process. The interviews will be semi-structured to leave room for follow-up questions and to support a deeper elucidation of participants responses and thinking processes[80].

1.5 Research Scope

In order to focus on the research objectives listed above, the scope of this thesis is limited as follows:

1. To create personalized browsing experience for the users, we have validated information extracted manually from Facebook. Other social networks like LinkedIn and Orkut are excluded from the discussion.

2. The focus of this thesis is on validating knowledge extracted from the Wikipedia pages for creating data-on-demand. We have not tested our system on other web documents like forums, and news articles.

3. To make the web documents, available in multiple languages, accessible for the user, our focus is only on creating different accessible versions of the page by using re-narration approach. Topics like quality control, system efficiency are not considered in this work.

4. In the thesis, we have utilized crowdsourcing approach, following fun and social incentives, to solve our research goals. Monetary incentive to motivate users for solving the given task is not explored in our research.

5. The thesis focus is on exploring ways to merge both automated and human computation to solve our research goals. This work is not aimed at maximizing system accuracy or machine performance. The main focus is on creating engaging experiences that could lead to higher participation of the users in the given task, which in turn will contribute to personalized web benefits.

1.6 Thesis Outline

The thesis is divided into four parts corresponding to three thesis problems, namely, extraction of user preferences, extraction of named entity and relations, and re-narrating the web documents. The roadmap of the thesis is shown in the Figure 1.3.

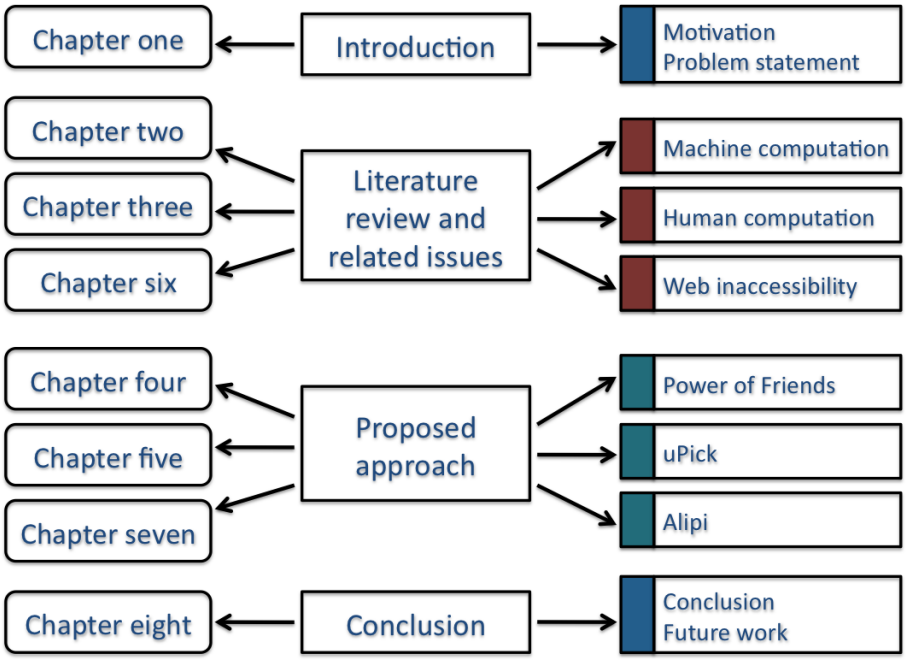

Figure 1.3 Roadmap of the thesis

Chapter 2, 3 and 6 address Objective 1 of Section 1.3.1. Chapter 2 presents an overview of how automated systems work to extract knowledge from the documents. The chapter explains different libraries popularly used to perform the task of extraction. The chapter concludes with the discussion of the need for human intervention to tackle the associated issues of the automated systems. Chapter 3 introduces human computational system and its different types. The chapter also describes three incentives commonly used to motivate users to perform a task. The later part of the chapter describes human computational games in detail and the problems related to designing such games. Chapter 6 discusses the existing works to solve accessibility issues of the web. We illustrate how these systems are unsuccessful in making the information available in an unknown lanaguage accessible to a user.

As a response to our Objective 2, Chapter 4 describes a novel Friendsourcing approach to extract user preferences from the community activity by proposing a human computational game called as Power of Friends; Chapter 5 describes our another crowdsourcing approach, uPick, to extract named entities and their relationships from the web documents; and Chapter 7 describes a crowdsourcing based re-narration framework, Alipi, to make the web pages accessible.

To fulfil our Objective 3, we analysed our proposed approaches by conducting user studies. Chapter 4, 5 and 7 also provide details of the study conducted to test our prototypes along with the study results and findings.

Finally, Chapter 8 concludes our thesis describing our research contributions and future directions.

1.7 Thesis Contributions

This research contributes to the understanding of crowdsourcing systems for personalized web both in practice as well as theory. This research makes the following contributions.

1. This research contributes to practice by providing implementation details and insights gained from the design and evaluation of three crowdsourcing based systems, namely, Power of Friends, uPick and Alipi, that demonstrate how collective human intelligence can support in the creation of personalized web experiences.

2. This work contributes to the understanding of interrelationship between the machine computation and human computation. By drawing key principles from crowdsourcing and HCI to support applications of the personalized web, the thesis aims to situate the role of collective human intelligence in solving web related issues.

3. This work expands the view of information retrieval systems beyond efficiency and accuracy by focusing on the task experience and invested efforts by the users.

Related Publications

- **Publication List**
  Following is the list of publications related to the research problems solved in the thesis. Much of the text in the thesis is taken from these publications.

  1. D. Aggarwal, R. A. Khot, and V. Choppella. Power of Friends: When Friends Guess About their Friends' Guess. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '13, Paris, France, 2013, ACM.
  2. D. Aggarwal, R. A. Khot, V. Varma, and V. Choppella. UPICK: Crowdsourcing Based Approach to Extract Relations Among Named Entities. In the Proceedings of *Proceedings of IndiaHCI*, Pune, India, 2012 (Accepted as full paper).
  3. T. B. Dinesh, S. Uskudrali, S. Sastry, D. Aggarwal, and V. Choppella. Alipi:A framework for re-narrating web pages. In *Proceedings of the International Cross- Disciplinary Conference on Web Accessibility*, W4A '12, pages 22:1-4, Lyon, France, 2012, ACM.

- **Public Demonstrations**

  1. Presented "Alipi: Making the web Inclusive and Accessible for All" in *IIIT-Hyderabad R&D Showcase*, Hyderabad, India, 2013.
  2. Presented "Crowdsourcing Based Approach to Extract Relations Among Named Entities" in *OpenData Camp-Hyderabad Meet*, Hyderabad, India, 2012.
  3. Poster presentations on "Power of Friends: Rethinking Games With a Purpose", and "Alipi: A renarration Web" in *IIIT-Hyderabad R&D Showcase*, Hyderabad, India, 2012.

- **Other Publications and Demonstrations**
  Following is the list of work that directly or indirectly helped me in writing the thesis.

  1. D. Aggarwal, H. Zade, and A. K. Dey. Demography based Automated Teller Machines. In the Proceedings of *IndiaHCI*, Pune, India, 2012 (Accepted as full paper).
  2. Poster presentation on "Umang: Emotions over-a-distance" in *MIT Media Lab Design Innovation Workshop*, Delhi, India, 2012.

3. Presented “Demography based Automated Teller Machines” in *International Conference Academic Showcase on Usable Software and Interface Design (USID)*, Auroville, India, 2011.
4. D. Aggarwal, J. Santerre, and A. K. Dey. Bayesian Analysis to improve Crowdsourcing platforms. In preparation.

Literature Survey

This page is intentionally left blank.

Chapter 2

Reviewing Systems with Machine Computation

2.1 Introduction

The area of Information Extraction [25, 42] strives to minimize efforts of the users by presenting them a holistic view of the available information on their selected topic. Information extraction refers to the automatic extraction of information from unstructured resources like web documents, blogs and discussion forums. These resources contain information of various entities, and relations describing them. The relations between the named entities are the facts about them and constitutes to the knowledge. However, the ambiguity (e.g., acronyms, abbreviations) and the errors (e.g., spelling mistakes) present in the available data make the task difficult.

The type of knowledge extracted from various resources falls under two categories [3, 97]:

Explicit Knowledge constitutes to those facts which have a definite answer. These facts are called as *Objective facts*. For example, 'It will rain tonight' is an objective fact because it might rain or it might not rain tonight. The truth value of the objective facts are accurately predicted and is independent of the human judgement. It can be represented in digital records and is always available for manipulation and representation.

Acquisition of explicit knowledge finds its applications in diverse areas, from commercial needs to scientific and enterprise applications. For example, automatic tracking of specific event types from news sources, building better search engine, question answering system and many more in commercial market as advertisements.

Tacit Knowledge constitutes facts which are subjective to individuals. It is restricted with certain context and is deeply affected by individual experience, aptitude, perceptions and inceptions. Because of the tacit knowledge, the interpretation of certain facts becomes subjective, and such facts are called as *Subjective Facts* or *Cultural beliefs*. Such knowledge is not expressed explicitly and resides in the minds of individuals, teams, and communities. Tacit knowledge is learned and shared to a certain extent by a group of people having common interests, similar perception, and ideologies [106]. These people form communities where members of the community know per-

sonalities, expertise, tastes, and perceptions of each other. Therefore, it is more appropriate to call tacit knowledge as Community Knowledge. For example, 'God exists in idols' is a cultural belief held by Hindu community. Since the amount of information in a culture is too large for any one individual to master, individuals know different subsets of the cultural knowledge and vary in their competence [106].

Community knowledge is a valuable asset and therefore, finds importance in various applications, such as, personalized marketing to personalize individual's computing and browsing experiences; prediction markets forecasting future events, like election polls; collecting subjective perceptions about the quality of a product; community advertising, such as advertising a Bengali concert to the members of a Bengali community; and to measure beliefs of people in non-measurable outcomes, such as whether the world will end or not. Moreover, getting to know about the community, its activities and varied opinions of its members also help in propagating relevant social messages. For example, the Facebook community of Anna Hazare's campaign[1] against corruption met a great success in bringing awareness among people.

Various techniques have evolved to perform the task of extraction with high precision. A large body of work[18, 22, 37, 46, 47, 54, 56, 67, 87, 92] now exists to tackle with different aspects of knowledge extraction and to address the needs of various applications. Early systems were rule-based with hand coded rules, which then gradually evolved to machine learning techniques. Every system has certain limitations either in its architecture, or implementation, or performance, or more than one together. Therefore, applying a system is based on the required application and the available data. We will discuss various techniques and their applications areas in detail in later sections of this chapter.

In the remaining chapter we will see how an extraction system works, and will list the existing techniques to extract both cultural and objective facts (Section 2.2). Section 2.3 describes limitations of automated techniques that make the task of knowledge extraction challenging and error prune. The chapter is then summarized in Section 2.4.

2.2 Knowledge Extraction systems

The task of knowledge extraction consists of three phases, each phase describing particular specifications of how a given task is accomplished. Figure 2.1 describes the working of an extraction system where input resources are given to the extraction system. The system pre-process the input and applies various methods to extract facts from the input data. The output generated is then validated to produce facts, and all the retrieved facts are stored in a database. Applications like search engine, question answering systems fetch records from the database. Below we describe each unit of a knowledge extraction system. A detailed survey of the working of extraction systems is present in the survey paper by Sarawagi[85].

[1] https://www.facebook.com/annahazare

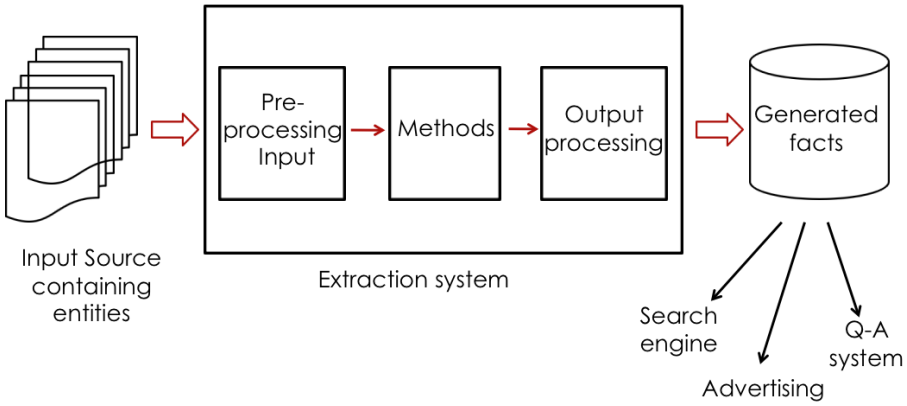

Figure 2.1 Architecture diagram of the Information Extraction system

2.2.1 Input resources

Input to the extraction system is mostly unstructured, and signify either sentence level small text or document level large text. For example, citations and classified ads are small text snippets, whereas extracting events from news articles and relations from a document are large text resources. Explicit knowledge can be extracted by applying extractors on such unstructured resources.

However, Community knowledge is never explicitly available in form of documents for analysis and most of the times it is scattered across the community network. Therefore, to extract this useful knowledge, sentence level parsing of social networks, community activities like public profiles, posts and comments, is required.

The implication of an extractor depends on the homogeneity of the given text. If the text follows similar style and format, for example, HTML pages generated via machines, a simple extractor will work with good accuracy. But if the format of the text is partially structured, such as, blogs, or news articles, then complex extractors are required to attain good accuracy.

2.2.2 Types of knowledge extracted from input resources

The standard types of structures extracted from the input resources are named entities and the relations between them.

Named Entities

An entity is a noun phrase forming an atomic element in a body of text. The most popular form of entities is formed by proper nouns, and is known as *Named Entity*. Common examples of the named entities are names of persons, locations, and organisations. Named Entity Recognition [25, 42] is one such subtask of the information extraction domain useful in gathering the factual information from the

given input text.

Relationships

Named entities when linked together, form a relation. This relationship generates the factual information required to grasp the meaning of each sentence present in the given text. Every relation is made up of three parts: Subject-relation-Object, where Subject and Object are the named entities belonging to the same or different types(Person, Location, Organisation) and a relationship between them is defined by a verb, an adjective, an adverb etc. The relations thus formulated, can then be used to present facts about a particular topic.

To understand the named entities and the associated relations, consider the following body of text:
Sachin Tendulkar was born in Bombay. His mother Rajni worked in the insurance industry, and his father Ramesh Tendulkar, a Marathi novelist, named Tendulkar after his favourite music director, Sachin Dev Burman.

In the first line of the above text, 'Sachin Tendulkar' is a named entity of type 'Person', while 'Bombay' is another named entity of type 'Location'. The relation 'born in' associates 'Sachin Tendulkar' and 'Bombay' together to form a relation: 'Sachin Tendulkar was born in Bombay'. Here, 'Sachin Tendulkar' is the subject, while 'Mumbai' is the object of the relation defined. All extracted relations from the above body of text are shown in Table 2.1.

Subject (Named Entity)	**Relation**	**Object (Named Entity)**
Sachin Tendulkar	born in	Bombay
Sachin Tendulkar	mother	Rajni
Rajni	worked in	Insurance company
Sachin Tendulkar	father	Ramesh Tendulkar
Ramesh Tendulkar	named	Tendulkar
Ramesh Tendulkar	favourite music director	Sachin Dev Burman
Tendulkar	named after	Sachin Dev Burman

Table 2.1 All possible relations from the above body of text

In the above body of text, Sachin Tendulkar is referred several times by different expressions. For example, 'His mother Rajni', refers to Sachin's mother expressed by using a pronoun 'his'. 'Ramesh Tendulkar, named Tendulkar' refers to Sachin Tendulkar expressed by using his last name. Referring the same named entity with various expressions by using the pronouns or adjectives, is called as *Co-reference*. Resolving the expressions by replacing them with the actual entity is called as *Co-reference Resolution*.

In the above piece of text, second sentence contains many binary relations about Sachin Tendulkar. Processing such sentences having multiple relationships require deeper analysis of not only each sentence, but also more than one sentence together. Therefore, pre-processing of the input data becomes essential to eliminate dependency of sentences on each other. Such processing is performed by using

Natural Language Processing(NLP) based techniques [74, 107] as co-reference resolution, discourse analysis etc. We discuss these techniques in next sub-section.

2.2.3 Pre-processing of input data

An extraction system is given the information of the types of structures to be extracted from the given input resource. Mostly, such systems do not take unstructured resources directly as input. Instead certain other additional resources are also provided to aid the extraction task. These additional resources include preprocessing libraries that enhance the given input with linguistic information. Such an enhancement with linguistic information serves as a valuable anchor for recognizing structures. Some of the popular and highly used NLP libraries are IBM's Languageware[2], libraries from the Stanford NLP group[3], and certain other mentioned as the OpenNLP effort[4]. These libraries are language dependent because they utilize the grammatical structure of a sentence to perform their tasks. Below are some pre-processing steps that are generally used by an extraction system.

Sentence tokenizer chunks the input document into individual tokens by analyzing the boundaries of each sentence in the given text. A token is typically a word, or a digit, or a punctuation, or a delimeter.

Part of Speech Tagger tags each token into a gramatical pre-defined category. For example, a sentence in English is tagged with tags like noun, verb, adverb, pronoun, article etc. An example of POS tagger on the above text snippet appears as follows:
Sachin/NNP Tendulkar/NNP was/VBD born/VBN in/IN Bombay/NNP

Parser utilizes Context Free grammar to identify the structure of a sentence in terms of its individual constituents. The output of a parser is a parse tree, which provides a better representation of how phrases are linked in a sentence.

Named Entity Recognizer (NER) identifies the named entities in the given text snippet along with their associated types as person, location, and organisation. For example, NER on the above text snippet will give the following output:
Sachin Tendulkar/PERSON was born in Bombay/LOCATION

Co-reference Resolution Tool resolves co-references at the document level and replaces all the co-references with the actual entity. Finding named entities and the associated co-references require extraction of the semantic knowledge from a large body of text, and therefore, is still a challenging extraction task.

[2]http://www.alphaworks.ibm.com/tech/lrw
[3]http://nlp.stanford.edu/software/
[4]http://opennlp.sourceforge.net

Word Sense Disambiguation refers to removing uncertainty of meaning from a word and provide it a specific sense intended in the given text. For example, a word 'crane' has different meanings in different contexts, as a noun it is a bird, and as a verb it means to lift or to move. The specific meaning of the word is determined from the textual context (surrounding few sentences) in which that ambiguous word has appeared. The automatic resolution of ambiguous words is handled by developing natural language processing techniques for specific domains like medical domain. Generally, the disambiguation of words is handled by using extrenal vocabularies as WordNet [76], DBpedia [2], MindNet [13, 92], or by building domain dependent vocabularies.

However, many of the pre-processing steps are expensive, and therefore, selective pre-processing of some parts of the given input text is considered to be more effective. Moreover, these libraries are trained on some specific data corpus like newswire, wikipedia, and for some specific domain. Therefore, these libraries have the tendency to introduce errors when used in some other corpus. If pre-processing step introduces errors in the system, then the later steps of extraction system are prone to be erroneous. For example, if a co-reference resolution tool replaces a co-referenced expression with a wrong entity, then the relations extracted for the replaced entity and the actual entity in later phase are also wrong. Therefore, to avoid cascading of errors, an extraction system needs to select these pre-processing libraries cautiously.

2.2.4 Methods of extraction

Various successful methods exist in the field of information extraction to correctly identify named entities from the corpus for different domains, e.g., medical, newswire domains. [37, 46, 56, 67]. Automated techniques that are used to extract named entities and the relations among them, can be classified based on the usage of the following methods: Natural Language Processing, Machine Learning, Statistical methods and some miscellaneous approaches.

Natural language processing techniques are rule based and are dependent upon the language structure of the input data. For example, a rule to extract entity relations from an English sentence can be 'noun-verb-noun'. Machine learning based approaches apply supervised and unsupervised learning of the system to train it for a given task. For example, to extract instances of some specific medical diseases from a given corpus, the system is trained with the symptoms of each given disease. Statistical methods are pattern based, and seek to detect some specific patterns from the given text to extract relations. For example, to find the birth date of a person from the given text, a pattern 'born in' is searched before a date. Here, 'date' may also be a pattern defined as DD/MM/YY. Furthermore, there are some systems that make use of some external vocabularies to extract relations. For example, maintaining a mapping file(vocabulary) of all the countries and their nationality, helps to determine the nationality of a person if his birth place is known.

However, these approaches are highly inter-linked. Therefore, a combination of these methods are generally used to perform an extraction task. For example, an extraction system can use NLP

based libraries to pre-process the input data, and a machine learning based approach to extract the relations along with some external vocabulary as WordNet. In the thesis, we have implemented rule based approaches, along with some pre-processing NLP libraries to extract named entity relationships from the online documents. These are few examples of the existing systems that are developed using CRF[67], rule based[18, 54], HMM[37], SVM[46], supervised learning[22, 87], MEM[56], context based clustering[47], and external vocabulary[92].

There have been various efforts to automate the process of community knowledge extraction. Since community knowledge consists of biased thinking, therefore, models which could incorporate such biasness are required. Moreover, the amount of knowledge inside a community is very large, therefore, individuals hold only a subset of this huge knowledge and vary in their competence. Automated techniques try to include all the human perceptions into the system to accomplish extraction task[51]. Cultural Consensus theory provides various models to incorporate cultural competence and perception of workers into the system as a variable[106]. Since it is hard to incorporate all the human factors into the system, therefore, extraction of community knowledge with automated techniques is always questionable.

2.2.5 Output generation

The extraction systems can be divided into three types, based on the type of output generated by them. First, a system which identifies all mentions of the strutured information, entities and their relations, from the given input data. For example, extracting all named entity relations from Wikipedia dataset. Second, a system which identifies missing relations of entities present in a given database. For example, given a database containing some entities and their relations, the extraction system extracts all the missing attributes of the given entities from the given corpus. Third, a system which link various entities of a database with some relationship. For example, given a SQL database containing named entities as table records, the extraction system links two entities together with some relation extracted from the given corpus.

The structures extracted, entities and their relations, from the input resource are stored in a database. In all the three types of extraction system, the extraction methods remain the same but the way of storing output differs. The database can be as basic as an SQL database where all the entities are stored as table records, or can be as much sophisticated as an RDF storage[5] where a relation is stored as a tuple(as subject-predicate-object expression). The choice of database normally depends upon the application or web-service, which is retrieving information from it.

Involvement of various natural language processing or machine learning based approaches in different phases of an extraction task tends to reduce the accuracy of the generated output. Often the results produced have high error rates. Therefore, many systems demand an additional validation phase to prune away candidates with low support. The validation step requires highly expensive statistical approaches and still can not produce 100% accurate results[32, 33].

[5] An RDF is a system to store and share data between computer programs.

2.3 Why do automated techniques fail?

Deployment of these extraction techniques to achieve an operational task raises many practical issues of performance, usability, and other engineering problems. We discuss them below.

a) Accuracy: Various techniques and tools [46, 56, 67] are developed to achieve higher accuracy in extraction task. The existing approaches can achieve 90% accuracy to extract basic named entities but for relationship extraction, the accuracy of the systems lie in the neighborhood of 50-70%. The reason behind such low accuracy is that each processing step invokes some issues and solving those issues introduces more errors in the system. Thus, errors get accumulated at each processing step resulting in an overall decrease of the system accuracy. Let us illustrate an example of cascading errors in an extraction system build on medical domain.

Word sense disambiguation of an ambiguous acronym 'RA' in medical domain has the following eight senses: 'rheumatoid arthritis', 'renal artery', 'right atrium', 'right atrial', 'refractory anemia', 'radioactive', 'right arm', 'rheumatic arthritis'. To disambiguate this acronym, extraction system seeks knowledge from other vocabularies such as WordNet [75, 87]. But such resources are liable to introduce words with different contexts(noun, verb, adjectives, adverbs) and make the extraction task tedious furthermore. The newly generated word along with its different context will generate wrong relations, thereby reducing the accuracy of the system.

Using various extraction techniques will, thus, allow unpredictable errors associated with the retrieved data and will lead to inaccurate information.

b) Dependency: Most of the extraction systems are developed for a particular domain (medical, protein) or a corpus (Wikipedia, web-blogs). In certain cases, the developed system aims to determine only some specific relations, such as extracting only DATE-and-PLACE-of-EVENT type of relations [22, 47]. Such systems, however, do not perform well when applied to other domains or corpora. Therefore, their usage remains limited to the domain or corpus for which they are designed. For example, a system developed to extract relations from Wikipedia Infobox can not extract accurate relations from web-blog dataset. The reason being change in the structures of Wikipedia and web-blog corpora.

A majority of such systems depend upon external sources or vocabularies. Such resources provide various semantic relations between different words, such as identifying hypernym, hyponym relationship between two words. These resources also provide automatic text analysis support for applications like automatic text classification and automatic text summarization. The issue with such dependencies is that these vocabularies are mostly maintained manually, and hence updating them is a costly task and also requires expertise. Moreover, they are limited in their size and are not specific to any domain leading to introduction of context related noise issues.

c) Efficiency: The pre-procssing steps and the feature detection steps are very expensive and are CPU-bound. Both rule-based and statistical approaches depend on feature detection to recognize en-

tities and their relationships. Features have varied cost of evaluation associated with them. For example, checking if a word is capitalized is cheaper, whereas checking the boundary of a named entity is expensive. Moreover, most of the systems require days of computation to extract relations from about a terabyte of data corpus.

2.4 Conclusion

The area of Information Extraction has various applications in different domains of medicines, search engine, marketing, analyzing data, and many more. In this chapter we discussed different aspects of knowledge extraction using automated techniques. Various systems are proposed to perform the task of extraction, which include ruled based approaches, machine learning based approaches, and natural language processing based approaches. However, the presence of ambiguity and noise in the available resources make the extraction task challenging. Current techniques are insufficient in handling scenarios where spelling errors and ambiguous acronyms introduce uncertainty in the meaning of a sentence. Moreover, solving such issues using vocabularies introduce errors in the system leading to accumulation of errors in every step of extraction. To avoid such cascading errors, extraction systems need to define each step carefully.

The existing systems have limitations in terms of accuracy, dependency, and their performance. Out of them, accuracy of the automated systems have always been a primary concern. State-of-the-art extraction models for basic named entities can achieve accuracy close to 90%. But the systems lag behind for relationships extraction, where even the sophisticated systems can achieve accuracy in the neighborhood of 70%, and that too, in a restricted domain. Most of the existing approaches are dependent on external vocabularies, or data corpus, or applications. Therefore, applying such systems to other domains or a different corpora remains an issue. Newer applications of Information Extraction technologies are increasing tremendously with the overwhelming generation of data on the web. Therefore, it is high time to investigate such approaches, which can solve the issues prevalent in the existing systems and are able to produce more accurate results.

Despite the promising results and improvements proposed by the existing systems, there remains a significant gap between the performance of such systems and the human intelligence. However, humans can not perform the whole task of extraction for millions of documents. Therefore, there is a need to discover and examine techniques, which can provide a blend of both the automatic techniques and human capabilities. The systems that incorporate human skills while performing a computational task are called as Human Computational Systems. We will discuss human computational system in next chapter.

This page is intentionally left blank.

Chapter 3

Reviewing Systems with Human Computation

3.1 Introduction

Before the dawn of the machine era, the term 'computer' refered to a person who performed calculation as her profession, and the machines that came later on, were named as 'AC' implying 'Automatic Computing' to distinguish machine computation with human computers [44]. In automatic computing, a human employs a computer or a bunch of computers to solve a particular problem. The human provides a formalized description of the problem to the computer and then receives the solution to interpret. As a result, most of the computations which were earlier processed by human computers can now be solved by using computers with very high effeciency. Such computations include calculation of correlation coefficients, regression analysis, differential equations, which required many hours of human computers labour, can now be solved in matter of seconds with computers. However, despite the superiority of computers over humans in doing rapid calculations and computations, there still remains many problems which humans can solve and surpass the performance of computers. These are the problems for which even the most sophisticated computers fail to perform effectively but humans can because of their higher cognitive abilities[82, 98, 99]. Such problems typically include problems related to Natural Language Processing(NLP) or Artificial Intelligence(AI) tasks, such as, image tagging, object finding, sentiment analysis, music classification. Moreover, there exists many NP-complete problems [72], such as, Travelling Salesman problem [43], which can not be solved within a reasonable computational cost by automated computers.

A new computational era thus evolved, where computers leverage the computation of a large number of human computers. Such system is called as *Human Computational system*, which tackles problems that neither computers nor humans can completely solve individually. In human computational system, the roles played by computers and humans are switched from that of traditional automatic computing. Here, instead of a human, a computer asks a group of people to solve a particular problem and then collects, interprets and integrates the collected solutions. Humans are involved in performing various basic operations, varying from basic computation operations (such as image labeling), to desiging control of the process, to sythesizing the output.

Solving a computational task in a human computational system entails three sub-units: distributing a problem to various workers, aggregating their outputs while rewarding them for their contributions, and then processing the collected data to final solution of the given task (Figure 3.1). A problem defined on human computational system is first decomposed into basic sub-tasks. The sub-tasks are distributed to many individuals in an organized way. Later, these individuals are given a set of instructions to perform the given sub-task. The output generated is then collected and validated to generate final output of the given problem. A human computational system follows an approach of distributing these sub-tasks to a closed set of workers recruited through a particular process of hiring. When the system hires human workers through an open call to distribute the task, the approach is known as *Crowdsourcing* [68].

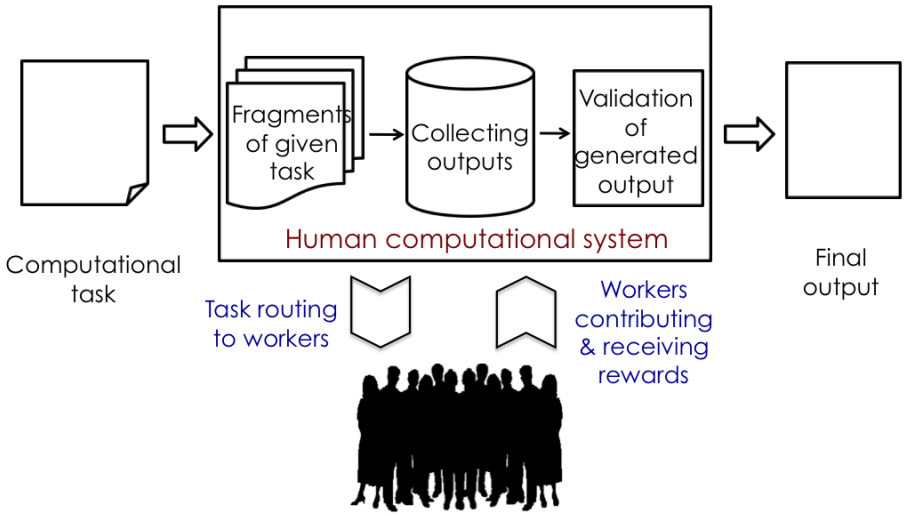

Figure 3.1 Architecture diagram of the Human Computational Systems

The human computational systems, thus, aim to solve problems by utilizing the collective intelligence that emerges from the collaboration of the recruited individuals. A major challenge to such systems is to build an effective algorithm that can route a task to those individuals who are capable of completing it efficiently. Various strategies have been proposed to allocate the tasks to capable, expert, interested and competent individuals. Examples of such strategies include questionnaire based score ranking, performance tracking of the workers etc.[51, 106]. However, none of the existing human computational systems is able to seek the confidence of a worker in performing the given task. Since the confidence is unknown, the system has to iterate over multiple rounds to collect data from several workers.

Human computational systems has two main end users: workers and requesters. Requesters are the people who distribute the task to a number of people, called as workers. To motivate people for their contribution, human computation system provides incentives to each contributor[57, 83, 88]. These incentives affect many aspects of a worker's contribution from her level of participation, to the quality

of her performance. To discourage workers from cheating the system, the system may also implement some punitive measures, such as revealing the worker identity as a cheat within her community. This could affect her reputation in her community and will discourage other requesters to assign her any task further.

Even though human computational systems provide various incentives to motivate people, the incentives may have their own drawbacks. These systems are still struggling to quantify what constitutes the appropriate incentive that motivates a worker producing high quality outputs instead of spamming the system[73, 109]. Incentives may be unable to motivate the users to produce better quality. Therefore, human computation systems require to implement some validation schemes to filter out the output generated by the workers.

In the remaining chapter we will discuss types of incentives that human computational systems provide to motivate the workers (Section 3.2). In Section 3.3 we will discuss different aspects of human computational games along with the issues in converting a task into a game. The chapter is then concluded in Section 3.4 stating the possible future work in this area.

3.2 Types of incentives provided by human computational system

Human computational systems require motivation schemes to invite workers. For some tasks, explicit incentives in the form of money, or virtual rewards (points, badges), are provided to the contributors. Such incentives come under *extrinsic motivation* scheme. While some workers have *intrinsic motivation* of fame, curosity, power etc. to contribute to a task.

Below we explain both types of motivation provided by the human computational systems. In our thesis, we have proposed approaches that make use of intrinsic incentives to improve the accuracy of automated systems, and to invite the social contributions. Therefore, in this chapter we focuss more on the challenges and issues related to intrinsic incentives.

3.2.1 Intinsic Motivation

According to Reiss et al. [83], humans are driven by sixteen basic desires, such as, power, curosity, status, honor, social contact etc. Individuals prioritize them differently with time. The strong and the weakest desire normally drives the behavior of a person. For example, some people read more to satiate their curosity, while others like to meet people and build good contacts. Each basic desire motivates individuals to perform some task, which can be differentiated into two categories: *Enjoyment based motivation* and *Community based motivation* [57]. Enjoyment based motivation contains factors that give a worker the sensation of "fun", whereas the Community based motivation covers those behaviors of workers that encourage her to be part of a community.

Enjoyment motivation is mostly carried out to kill time and to achieve a sense of accomplishment. Games are designed to give workers fun along with direct feedback of their performance. Such games

that are designed to realize tasks from both of the above mentioned categories, are called as *Human Computational Games*. We will discuss these in detail in Section 3.3. Community based motivation is guided by personal identification and building social contacts with peer community members. Example of community based motivation system are: Wikipedia[1], discussion forums of Mozilla[2], and Javascript[3].

Intrinsic motivation gives full autonomy to the worker as she is allowed to perform the task whenever she wishes. Intrinsic motivation can produce better quality of results. This, however, is not enough to indulge a larger crowd for their contributions to a given task. Because such motivation requires familiarity with the task and some pre-defined skills set to perform it, and thus it targets at particular groups only. For example, to participate in a Javascript discussion group, a person needs to be skilled in Javascript. Moreover, the motivation of such skilled contributors is generally temporary and lasts for a duration of time required to fix a bug in their programs[64, 79]. Wu et al.[108] studied the factors affecting contributors of YouTube, and demonstrated that contributors who lack in receiving attention tend to stop contributing, while prolific contributors attracts a large crowd and keep on uploading videos on YouTube. It has been observed that such motivation is highly intermittent and is dependent on various factors such as learning, attention, altruism, competency[50, 79]. Therefore, building human computational systems based on intrinsic motivation in such a way that interest of the contributors is sustained, remains a challenge.

3.2.2 Extrinsic Motivation

Extrinsic motivation guides a worker to perform an activity to achieve a certain desired outcome in the form of money, or social rewards and badges. We explain both of the categories as follows:

- *Monetary Incentives*
  In monetary incentives, a worker gets paid for his contribution. The best example of such a system is Amazon Mechanical Turk (AMT) [1]. AMT is a financial market place for small tasks such as: labeling images, annotating Named Entities and spam identification etc., where workers get some money for a defined unit of task [109].

  Studies have shown that money is not always an incentive to generate quality results. The study performed by Mason and Watts [73] concluded that the increased monetary incentives only increased the quantity of the performed work but not its quality. It is believed that larger financial incentives lead to more problems associated with validating the task done by the users along with other issues of dissatisfaction [60, 62]. Moreover, the system also faces issues of dissatisfaction bewteen workers and requesters. There are cases where workers demand more money for their contributions, while requesters do not find their work useful to consider. Therefore, monetary incentives lag to motivate workers to produce truthful responses.

[1] http://en.wikipedia.org/
[2] http://support.mozilla.org/en-US/forums/
[3] http://www.hotscripts.com/listing/javascript-discussion-forum/

- *Social Rewards as incentives*

  Giving social rewards is the extrinsic counterpart of the intrinsic motivation where the worker tries to establish her personal identification and social contacts. Unlike the intrinsic motivation of establishing social contacts, social rewards extend its significance to be identified outside the community and seeks feedback by other individuals like requesters. *Gamification* is one such technique that has evolved to engage workers in non-gaming applications by providing social rewards[29]. Few examples of the non-gaming applications are completing surveys, physical exercise and reading online documents. Gamification enhances such tasks by providing game mechanisms and game thinking, so as to encourage users to accomplish these non-engaging task. For example, Green Button[4] encourages people to save electricity by giving them social badges. The game builds a social environment by connecting many people and provides scores on reducing power consumption.

  Unlike the intrinsic motivation, incentivization through social rewards are highly governed by the provided competitive environment and performance outcome. Since the rewards of workers are based on their performance, the actions of workers are obligated by different rules and therefore, they tends to feel controlled. It has been observed that social rewards tend to weaken the intrinsic motivation of a worker to perform the given task over a period of time[48, 57]. Chasing rewards sometimes demotivates workers and has a negative impact on the creativity of the workers when they are unable to succeed.

The studies of behavioral psychologists show that extrinsic rewards do not undermine intrinsic motivation [93, 23]. They argued that it is the poor operationalization of the rewards to achieve short-term goals that could have negative effects on the workers. Therefore, extrinsic motivations can be utilized to foster intrinsic motivations, provided the task is designed in a positive manner to improve the ability of workers. The presence of extrinsic incentives should catalyze the workers motivation intrinsically.

3.3 Human Computational Games

Human Computational Games project a problem in an entertaining game-like environment where fun, curiosity and intellectual challenges in the games ensure enriched user experience [82, 99]. Such games are also called as *Games with a purpose(GWAP)* [5]. But the purpose of the game is normally kept hidden from the players. People play such games mostly for the sake of entertainment and accomplish the underlying tasks unknowingly. The first and the best-known example of human computational game is ESP Game[98]. ESP game is designed to tag images in order to produce better image indices(hidden task). These indices are then used to improve search results of retrieving images.

Human computational games work on the assumption that if large number of online players play these games regularly, then they can generate a huge amount of useful data collectively. This assump-

[4] http://news.thomasnet.com/green_clean/2012/04/10/update-on-green-gamification-could-the-green-button-be-the-killer-app-for-consumer-energy/

tion, however, holds true only if the game is entertaining and fun to play. Therefore, it is essential to look into game design properties that make games more interesting and fun to play.

Existing human computational games are generally designed using one of the following three game mechanisms[99]:

Output Agreement: In the games based on output agreement, two randomly paired players are given the same input and they try to agree upon the output valid for the given input. Players are not allowed to communicate with each other and they get to see each other's outputs only if there is a match. Therefore, the best strategy for players to win and to reach an agreement is to play honestly, and provide output related to the common input. ESP game[98] falls in output agreement category where players either generate or agree upon a valid tag for a given image.

Input Agreement: In this cateogry, two randomly paired players try to agree upon whether they received the same or different input. Each player produces output, which describes her input and can be seen by the other player. Both players after accessing each other's outputs decide whether they received the same input or not. On proper assessment about the input (same or different), they get rewards (points). A popular game based on input agreement is TagATune[69] that allows annontation of music and sound to improve search of audio files.

Function Computation (Inversion Problem): Unlike Input and output agreement based games, Function computational games are asymmetric and unidirectional, that is only one player is actively involved in contributing to the task at a time. In a typical implementation one player takes up the role of describer and helps the other player (guesser) in computing the function by giving relevant hints. Describer, therefore, can score once she passes correct information to the guesser such that the guesser is able to identify the given function correctly. Some examples of such games are: Peekaboom[102] to improve computer vision algorithms by recognising objects in an image, Phetch[100] to improve image search by labelling images with descriptive tags, and Verbosity[101] to develop intelligent computer programs by collecting common-sense data.

3.3.1 Games for building Web 3.0

Web 3.0 is about weaving web where all the data is linked together by some meaning. This linked data as discussed in Chapter 1 forms what is called as Semantic Web. The evolution of Web 3.0 is realized by the human contributors, playing different roles of data creator, extractor, validator, and consumer. One of the characteristics of Web 3.0 is the evolution of communities for almost each and every topic. Human computational games are also explored to generate sources of Web 3.0 in various ways. We discuss some of the games developed to weave Semantic Web and to extract community knowledge.

Recently, Siorpaes et al.[89] have reviewed multiple human computational games with the focus on the development of Semantic Web[45], where people contribute to weave the Web with a meaningful

linked structure. The idea is to bring human intelligence as part of a game in building such a formal knowledge structure of Web, which otherwise, cannot be fully automated for certain tasks. Some examples of such tasks include collecting named entities, finding relational hierarchy, phrase detection, finding neighbors in graphs and many others. Some examples of such games are Peekaboom[102], Verbosity[101], TagATune[69], with the computational purpose to collect database of image related tags, commonsense facts about words and music, respectively.

Some of the examples of Human computational games developed earlier to extract community knowledge include Collabio[19], 21 Questions[5] and Dogear game[35]. Collabio and 21 questions provide a social tagging environment for friends to tag relevant information about each other, while Dogear is a game embedded inside an enterprise social bookmarking system. In our thesis, we have also developed a novel human computational game called as Power of Friends to extract community knowledge. The details of the game is presented in the Chapter 4.

However, the system suffers from a variety of problems when one attempts to extract the required data from the underlying task of human computational games. We mention below few of such problems related to developing human computational games.

3.3.2 Issues with Human Computational Games

Existing human computational games suffer with various problems, which make them uninteresting for the players. Games built around the task of extraction, such as Peekaboom[102], Verbosity[101], TagATune[69], proved to be task oriented and therefore, are not able to entice the players for a long time. The design of such games, however, pose a variety of challlenges when one attempts to extract useful knowledge from them. We discuss some of the issues associated with the existing human computational games as follows:

- *Lack of challenges (Randomness)*
  Most of the existing human computational games have monotonous gameplay. Even though the computational task is hidden from the player, she knows that she can score high in the games only if she performs the underlying task well. For example, in ESP game, both players know that if they tag the image with easy and obvious tags, they can score good points. Although the knowledge of game objective(what to do in a game) and the winning strategy(how to score high) are beneficial for players, even then they tend to loose their interests in such games over a period of time. The game becomes redundant and boring, thus contributions from players tend to diminish quickly over time. A good game design must continuously provide wide range of challenges and winning a game should not appear easy, that is the game strategy and its objective should not be obvious to the player[65]. Moreover, the players should feel and experience a gradual increase in the satisfaction and learning which entice them with the game.

[5] http://apps.facebook.com/twentyoneq/

- *Problems with random pairing of players*
  Social interaction is one of the biggest factor of a game design that motivates a player to play the game even if she does like that particular game or games at all[91]. Most human computational games, therefore, incorporate a collaborative[98] and competitive[61] environment, where two or more randomly selected players play the game together and agree on a particular task in hand. But multi-player gaming model restricts a single player from playing when there is an absence of other willing partners. To mitigate this problem, Siorpaes et al.[89] has proposed an idea of a single player OntoBay game where the player plays against the previous game challenges and the past user inputs.

  However, we find two drawbacks in random pairing of players. First, it is difficult to design a game that ensure an equal satisfaction to randomly paired players, owing to their different levels of expertise. This seriously limits the human computational game design to only common tasks, such as, image tagging. Second, random pairing may encourage foul play and cheating in a game since identity of players is anonymous and not tied to their profile. A better game design is to follow community based design and to build games for communities, such as, games developed for friends. Studies show that players want to play more with their friends rather than a complete stranger[89]. Another possible direction is to design single player version of the game that employs asynchronous gameplay.

- *Clear articulation of the cognitive activities*
  Another challenge with human computational games is the clear articulation of the cognitive activities behind the games. For example, some of the human computational games require high cognitive efforts such as: writing descriptive tags, lengthy task of watching, listening the participants, thereby, making the game uninteresting for the users to get involved. Games built on solving NP-complete problems, such as, satisfiability problems[20], and scheduling and mapping problems[70] are examples of such games. We believe that the games should involve minimum cognitive efforts from the participants so that the hidden computational task of the games is not exposed to the users, thereby, increasing the entertainment value of it. Moreover, games should also be intellectually stimulating along with their rich fun quotient.

- *Motivating workers to generate accurate outputs*
  While designing a human computational game, a game designer has to look for both correctness of the underlying task as well as the entertainment quotient of the game. Balancing them together proves to be extremely difficult. Studies have shown that the games, which give high entertainment value often, produce a lot of noisy data [69]. On the other hand, task-oriented games are less popular among the users since they lack in the fun element [82, 99, 101, 102].

3.4 Conclusion

Since last few decades, we are witnessing a change in the way, we perceive and approach computing. This perception of computing has been altered with the dawn of human computation. A human computational system is defined as a system that designs a solution to a computationally hard problem, using both automated and human computers. Ever since its inception in 2005, human computational systems have been widely experimented both academically and in the industry, and were quite successful in solving variety of different computationally hard problems. These computational problems often require higher congnitive efforts from workers end, and thus, require incentives to motivate workers for their high quality contributions.

The type of incentive given to the workers depends on the task to be computed. If the task tends to be cognitively high, then extrinsic incentives are required. Human computational games have been very promising with respect to generation of quality results when designed properly. Maintaining the trade-off between fun element and the data collection is a thoughtful, though necessary, aspect of game design. Proper articulation of task is required to enhance the ability of workers, rather than leaving a negative impact on their ability and creativity.

The existing human computational systems are unable to identify efficient workers to accomplish a task successfully. There are cases where mallicious workers try to spam the system or cheat the system to get extra benefits either in terms of money or something else. Therefore, requesters also need to implement various cheating detection mechanisms to discourage such workers. This increases complexity of the system resulting in multiple iterations to collect huge data, out of which, a large quantity is noisy data and thus discarded. This necessiates to develop an efficient task routing algorithm, which can identify efficient workers and produce quality data. Measuring confidence of the worker in her contribution may be one aspect, which helps not only to reduce the number of iterations, but also to properly judge the value of her contribution.

A good design of human computational system provides a wide range of challenges to the players continuously. The task completion should not appear easy and obvious to the worker. Moreover, the workers should feel and experience a gradual increase in the satisfaction and learning which entice them with the task.

This page is intentionally left blank.

Problem 1- Extraction of user preferences

This page is intentionally left blank.

Chapter 4

Power of Friends: Friendsourcing Based Approach to Extract and Validate Cultural Beliefs

4.1 Introduction

Communities are the backbone of a society. Communities are formed by people having common interests, similar perception, and ideologies. Such communities represent a perfect platform for its members to come together to share their ideas and discuss common issues. Members of the closely connected communities often contain a wealth of knowledge about each other as well as about the community. This knowledge is known as Community Knowledge. Community knowledge consists of the beliefs related to either the community members or the community itself. These beliefs are learned within the community and are shared to certain extent outside the community.

The extraction of cultural beliefs finds its importance for various applications, such as, prediction markets, which are used to forecast future events (For example, result of the election polls), certain subjective perceptions about the quality of a product, and to gather beliefs of people on unmeasurable outcomes (For example, questions like whether the world will end). Such extraction is also useful in recommendation systems and for targeted marketing of a product to a particular group of individuals based on their beliefs. Moreover, cultural belief extraction can be useful in maintaining communal harmony and extracting and propagating certain relevant messages within the community members.

However, extraction of community knowledge is challenging. The community knowledge is ingrained and scattered within the community, and therefore, is never available explicitly for extraction. Several approaches have been proposed to automate this process of extraction from such communities while keeping into account various factors such as biasness and expertise of participants, data imperfection and timely changes in the cultural beliefs of people. However, automated techniques fail to model different perceptions of people into a working system. In Chapter 2, we have discussed that automated approaches include extracting beliefs through statistical methods, machine learning and Natural language processing based systems. However, despite the advancements in machine learning techniques and artificial intelligence algorithms, the level of accuracy that humans can achieve in generating such

useful knowledge still remains unmatched. Therefore, it becomes necessary to involve humans for accomplishing the task of extracting community knowledge.

In the recent past, human computational games (games with the purpose) have emerged as a popular method of engaging participants in otherwise banal tasks such as the one mentioned above. The assumption here is that people will play these games for fun and entertainment and will generate useful data in return [68, 99]. When these games target close group of people (friends), they are called as Friendsourcing games[19]. Friendsourcing is a community based Crowdsourcing approach that motivates individual members of the community to share and validate accurate information about each other. The approach leverages game play elements such as competition and social accountability to entice the members in the activity. The Friendsourcing platform invites perceptive thinking of the individuals about her fellow community members. Based on Friendsourcing, the game Collabio generates tag clouds containing relevant information about the individuals[19]. Collabio succeeds in fetching information about an individual from her social network community. Some other examples of the existing Human computational games developed to extract community knowledge include 21 Questions[1] and Dogear game[35]. 21 Questions is a Facebook game with fixed set of questions. The game became popular as people find it intersting to answer questions related to their friends and to find out their response in return. Dogear game is about sharing bookmarks in a large network of people and recommending websites and documents of interest to each other. The popularity of all these games show that people are interested to tell about each other and to know what others are saying about them.

The existing games, however, suffer from the problem of noisy data and hence require extensive post-processing to filter out the proper tags[19]. For example, removing spelling errors, handling ambiguity of the generated tags, and other NLP related problems[68, 89]. The reason behind noise generation is that the data generated by these games remain at the level of lexical resources, that is, the generated terms and tags are not connected semantically. Moreover, these games struggle in balancing the entertainment quotient of the game with the correctness of the desired output[66]. Therefore, the popularity of these games has decreased over time. For example, Collabio game on Facebook has failed to attract followers and therefore, now has only 98 active monthly users[66].

In this chapter, we present a new mechanism called as *Crowd Consensus* for the human computational games to perform the task of community knowledge extraction and validation. Crowd consensus incorporates randomness in performing the task, which makes it more challenging and entertaining. It motivates people to contribute to the computational task of knowledge extraction to the best of their abilities while ensuring the correctness of the underlying task. As a proof of concept, we present an entertaining asynchronous competitive game called as *Power of Friends*. Power of Friends is a community based game for generating and validating facts about friends in a social network. It is based on popular TV game show Power of 10. We describe the complete design of the Power of Friends game and report the results of the empirical studies. Empirical studies of Power of friends have been extremely positive. We found that our approach is able to elicit truthful responses and the participants enjoyed our approach.

[1] http://apps.facebook.com/twentyoneq/

Users found the scoring and user interaction elements of the game highly enjoyable and most players wanted to continue playing.

The rest of the chapter is organised as follows: Section 4.2 describes the motivation of our work. Section 4.3 describes our proposed approach, Crowd Consensus, to Friendsourcing games. Section 4.3.1 gives a detail of the pilot study conducted to investigate our proposed approach. Section 4.3.2 motivates the need to propose a game design based on the proposed approach. Section 4.4 describes the proposed game design with gameplay explained in Section 4.4.1. Processing of responses and scoring a player is explained in Section 4.4.2. Section 4.4.4 contains the user study performed to evaluate the proposed game. Lastly, we conclude the chapter in Section 4.4.8.

4.2 Motivation

Humans being social animals like to interact, socialize and get opinion of others throughout their life. We live in a social world and our interactions, actions, thoughts are guided and influenced by people around us. This theory is devised by Charles Horton Cooley in 1902 and is known as *Looking-glass self*[28]. He describes an individual as, "I am not what I think I am and I am not what you think I am; I am what I think that you think I am"[12]. Cooley argues that the growth of an individual depends upon his interaction and relations with surrounding people. He suggests that the perception of others about the individual often shape up his personality. Therefore, we believe that it is essential to obtain the opinion of community members while generating information about an individual.

Additionally, *Cultural Consensus theory* suggests that the amount of information inside a community is quite large for an individual member to master, therefore, every individual knows a different subset of it[106]. On the rubrics of this theory, we believe that to reach an agreement about a belief within a community, it becomes essential to involve a large pool of community members in extracting and validating the community related information. For example, we need to ask the community members a set of questions related to their community and other fellow members; and then validate the answers by aggregating all the responses.

However, revealing truth about a community member may involve element of social awkwardness where a member is not comfortable to give her response about her peer member because the truth may influence her relationship with the member. For example, if a student is asked to rate her teacher in front of her class teacher, chances that the student will rate her teacher as good is very high. While, if the student identity is kept anonymous, there are chances that she may give fair rating to her teacher. Therefore, it is required that responses of the community members are kept anonym while collecting their inputs. This problem is referred as *Secure multi-party computation*[34], where different parties jointly compute a function while keeping inputs of every party hidden from others. Therefore, we also believe that to generate truthful and unbiased community beliefs, every input from community member should be kept private.

Relying on the rubrics of the above mentioned concepts, we built a framework called *Crowd Consensus* as discussed in next section. The framework collects the perception of community members about the preferences of an individual. Crowd Consensus is developed to extract and validate community knowledge from the associated members of a community. Our framework is based on the premise that people belonging to the same community, over a period of time, tend to think alike[26, 52]. Therefore, we believe that one person can accurately predict the opinion of her fellow community members. It may also happen that we need not to ask the opinion of every member of a community about a particular belief; instead, few members would be able to provide opinion of fellow members. We explain our proposed approach in the next section along with the studies conducted to test its feasibility.

4.3 Crowd Consensus: Rethinking Friendsourcing

There are two ways of validating particular information about an individual, let us say Meera, using Friendsourcing. The first and the simplest method is to ask Meera directly. However, Meera may not be voluntarily interested in disclosing information about her. The second method is to friendsource this task where information related to Meera is asked and validated by asking her friends. We find two problems with the existing friendsourcing approaches. The first problem is that despite the embedded game based narrative, the task of generating information by answering questions about an individual is repetitive and hence becomes boring and non-engaging after a period of time. We believe that a good game always provides a wide range of challenges to the players and the game strategy is not obvious to everyone [65, 66, 99]. In the existing friendsourcing games, the game challenge remains same every time and thus makes the game experience monotonous. The second problem is of discomfort to reveal information about a group member. Her friends may not be comfortable in disclosing certain personal information about her. The reason being that doing so may influence their friendship. One way to solve this problem is to mask the identities of friends willing to respond to the questions. However, anonymous identity may tempt friends to contribute fake responses resulting in irrelevant and noisy data generation.

As a solution to the above-mentioned problems for generating useful information about group members, we introduce a new approach to friendsourcing called as *Crowd Consensus*. The approach involves guessing about friends' opinion about some information. Unlike the existing approaches, our approach does not involve asking direct questions from an individual about some information related to her friend. Instead we reframe this monotonous question answering as a challenge and ask her to identify what her friends might have answered for the same question (Figure 4.1). For example, in order to know whether a group member Meera likes swimming, we ask, "How many of your friends think that Meera likes swimming?" in place of a mundane question: "Do you think that Meera likes swimming?" Although the individual may not be able to predict the opinion of her group correctly, we believe that this added guesswork will entice her to challenge herself and will prompt her to play more. Moreover, since the individual predicts the collective opinion of her friends about a particular question, she does not explicitly

mention the identity of any member of her group. This strategy of speaking about the whole group rather than a single individual will not involve any discomfort while answering and hence should encourage participants to reveal more truthful responses. The proposed approach also serves personal benefits for the members such as to identify how well their peers know them as well as what is the perception of peers about them.

We refer likings, preferences and interests of individuals in the community as community beliefs. For each belief, opinion of every community member may differ. Our approach is to gather opinion of the community members about a given belief. Therefore, we define a belief as a piece of information whose truth value within a community is unknown and every individual may have a different opinion about that belief. Collective opinion of all the community members will define the truth value of the belief. These beliefs are extracted from different activities of the community members and are related to either the community itself or to its associated members.

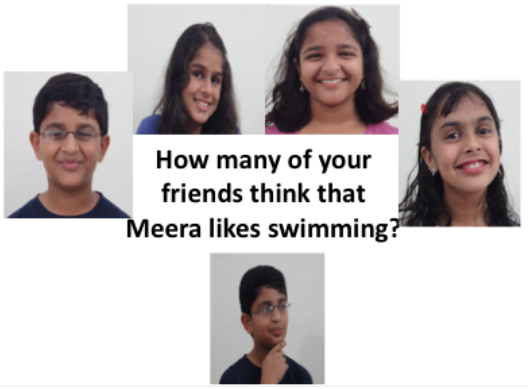

Figure 4.1 Approach followed by an individual to answer the given question: He polls for the opinion of his fellow members by keeping them in mind. He then gives a rough guess of how many other members have similar thinking.

For easy understanding, let us illustrate our approach with an example. For a belief "Raj likes Football", let us define three values as A, C and G. These values describe the ways of determining the truth value of the given fact. Table 4.1 describes the meaning of these three values. We look forward to determine the value of G using our proposed approach. We believe that the value of G is in agreement with the value of A as well as C.

4.3.1 Pilot Study

In order to investigate our proposed approach, we conducted a paper pen based supervised pilot study to test the following questions:

Q1: Does individual opinion about her community opinion is in agreement with the opinion of her community?

Truth value	Description	How to collect the value with example
A	The actual truth value of the given belief.	Truth value of the belief "Raj likes Football" is collected by asking it from Raj itself.
C	The community opinion on the truth value of the given belief.	Community opinion about the belief "Raj likes Football" is inferred by aggregating the responses of all the members of his community.
G	Individual's opinion on her community opinion on the truth value of the given fact.	Community opinion of individual members on the belief "Raj likes Football" is collected by individual's response on the belief (via our approach).

Table 4.1 Description of different terminologies to determine truth value of a belief

Q2: If answer to Q1 is yes, then what is the level of accuracy by which an individual can predict the opinion of her fellow members? Also, what are the external factors influencing accuracy of the underlying task?

Q3: Does the community opinion reflect the truth value of the given belief?

Using the terminologies defined in Table 4.1, the questions Q1 and Q3 can be formulated as whether G==C or C==A respectively are true or not.

Participants

We selected 32 participants (20 female) and formed four communities among them for our study. We named them as P1, P2, P3 and P4 (Table 4.2). The average age of the participants was 24 years with 22 years as the minimum age and 28 years as the maximum age at the time of study. All the communities were selected from our university campus. Most of the participants were from Computer Science engineering background. We provided chocolates to each participant for their contribution.

Id.	Type of community	Community age	No. of members
P1	MS batch 2010	2	8
P2	M.Tech. batch 2011	1	8
P3	M.Tech. batch 2010	2	8
P4	Professionals	2	8

Table 4.2 Description of the selected communities

Setup

We developed four paper based forms for every community each having 6 questions. In the form, we explicitly mentioned the names of people forming a community. Participants were told to give their reponses considering the community we have formulated. Please refer Appendix A.1 for a sample

paper prototype for one of our formulated communities. We asked each participant to fill the following information for each given belief: their individual opinion in yes/no, their opinion on the crowd opinion in yes/no and a rough estimate of the number of other fellow members who agree for the given fact. The prototype contained beliefs for the selected communities. All the beliefs were extracted from activities of the participants on the social networking website, Facebook and our personal interactions with them. Almost all the selected participants are connected through Facebook, therefore, we prefered Facebook over other social networking websites.

We aimed to analyze different factors influencing the accuracy of the collected data from our approach. Therefore, we introduced several factors such as community age, community size, types of community and types of questions into our study. The questions were related to entertainment(songs and movies), technical skills, hobbies(drawing, sketching, playing instruments, art skills), food habits and dressing style. These varying factors gave us insights on how much people were aware of their peer-group preferences and how much they were motivated to know more by our game. We specifically did not mention any personal question related to anyone. Although our institute does not require IRB approval to conduct user study. We still followed the standard practices. The demographic information of the participants were not revealed anywhere.

Procedure

The study was organized in two phases. The first phase of the study was dedicated to registration where demographic information of the participants was taken. We also asked the participants about the type of discussions they are involved with other community members and the bonding level among them. The session started with a brief explanation of the task. We explained the participants how to approach the given questions while predicting thinking of their community members. Each participant was then asked to fill the paper prototype of our approach related to her community. At the end, we collected qualitative feedback from the participants through open interviews. To know experience of the participants with the given task, we encouraged them to think aloud while filling the given form.

Results and Findings

For each belief, we collect majority votes for individual thinking and individual opinion about her community thinking. Individual opinions of all the community members formulate the community thinking, defined as C. The truth value of a given belief is collected from the individual opinion of the person with whom the belief is related which is value A. Individual thinking about the community opinion contribute to value G. We present the results of the pilot study in the Table 4.3. Column 2 and column 3 of the table respectively explains our research questions Q1 and Q3.

We found that for more than 90% of the beliefs, there is no conflict between the individual thinking and community thinking. We consider those beliefs as incorrect which could not obtain majority votes from the community memebers. However, the accuracy of community opinion reflecting the truth value of the given fact is 50%. The reason behind this low accuracy is unawareness of the community

Community Id	For how many beliefs individual opinion about her community opinion is in agreement with the opinion of her community (G==C)?	For how many beliefs, community opinion reflect the truth value of the given belief (C==A)?
P1	5/6	5/6
P2	6/6	4/6
P3	6/6	4/6
P4	4/6	3/6

Table 4.3 Results of the pilot study: Column 2 and column 3 are answers to our research questions Q1 and Q3 respectively.

members about the preferences of their fellow member with whom the belief was related. Community opinion is developed on how an individual portray herself to her community. Moreover, it also depends on how much people interact with each other. The bonding of a person with her fellow community members decides how much she knows the preferences of other members of her community. For example, the members of community P4 rarely talks about personal preferences with each other. Community members of P4 found it surprising that they never talked about certain general preferences of their fellow members. Therefore, *the most important factor influencing accuracy of the collected beliefs is the closeness of people with their fellow community members*.

We observed that *the highly motivating factor for an individual to play the game was the number of questions related to her*. If the number of questions related to her is more, she found the game more entertaining. Moreover, if there is not even a single question related to her, she lacks motivation to play the game. This interesting finding came out through our paper based form where we had six questions for each community. Therefore, beliefs related to any two members of each community were not mentioned. On asking the reason behind such a behavior, one member of P1 mentioned, "These questions are giving me the feeling that I make some difference in the group and my friends are aware of my likings." Another participant of P3 said, "If others have some likings and habits then so do I. I am also willing to know how well my friends know me." They are very much interested to know what others think about them. Therefore, it is essential to make people realize that they are involved in the game.

The other affecting factor was the type of questions asked from the participants. Since we were willing to see what types of questions are fun to play, we included different types of questions in our game. We found that people were not interested to play with questions related to technical skills and found such questions boring. One participant said, "I am a geek and I know everyone knows this aspect of mine. I am more interested to see if my friends know my habits and interests or not." Few members of P1 found the questions very generic and obvious to play with little guess work. Members of P1, P3 wanted to play with more personal questions so as to tease each other. The most unexpected finding was that participants were very much willing to generate questions about themselves and related to their

friends. The motivation behind such an enthusiasm was that they wanted to test their friends knowledge and to create more fun in the game by asking personal questions.

Participants liked the guess work involved in performing the underlying task. One of the participants said, "I did not know some of the questions about my friends but thinking of their personality, I could give a sure shot guess." Most challenging for them was predicting exact number of their fellow members who agree for the given belief. Answering in boolean(yes/no) was easy for them. But predicting the number required a lot of thinking. However, participants appreciated this part of the task and found it helpful to relate themselves with their fellows.

4.3.2 Discussion

Results of the pilot study show that our approach has good potential for extracting community knowledge with good accuracy. Results of the pilot study shows that individuals can accurately predict the opinion of their fellow members. However, the accuracy of the collected beliefs differed for different communities. For 90% of the beliefs, there was no conflict between the individual thinking and the community thinking. Whereas, only 50% of the beliefs were identified correctly by the community. The reason why individual thinking and crowd thinking at times, do not match with the truth value of the belief is lesser personal relationship among the community members.The closeness of people with other community members affects the accuracy tremendously.

In our study, we extracted questions from Facebook which were very general. We found that the game should include personal questions which will increase fun and also will help community members to know each other better. If members of a group are closely related with each other then one person can determine thinking of other members accurately. The most positive finding to our game was that people were very much enthusiastic to generate questions about their community members. No other social networking game so far developed has this feature where people can contribute questions about each other. All the games are dependent either on the automated techniques or manual work to extract preferences of community members. Moreover, the advantage of generating questions by asking community members itself is that the beliefs will be more accurate than that of the automatically extracted ones. In addition, the overhead of generating questions is also taken care by the game players.

However, the paper based approach that we have used to perform this task is not scalable. We need an entertaining and engaging automatic approach where people can participate frequently to generate more community beliefs. Human computational games come to rescue where we can involve larger crowd to produce bulk information about community and its associated members.

Designing human computational games is a challenging task and we have mentioned various related issues in Section 3.3.2 of the same chapter. We believe that the attributes mentioned in the section such as randomness, providing multi-player environment, lesser cognitive tasks etc. are essential for an enriched game experience and should be incorporated into a human computational game design. However, we should also take care that the accuracy of the underlying task is not compromised in doing so. Moreover, the players should experience a gradual increase in the satisfaction and the learning levels,

which will entice them to play more[65]. These attributes have motivated us to develop the following novel game design.

4.4 Proposed game: Power of Friends

Based on our proposed approach Crowd Consensus, we developed a human computational game called *Power of Friends*. Power of friends is a simple yet entertaining game for extracting and validating useful information from the community members using Friendsourcing. A snapshot of the game design is presented in Figure 4.2. We take inspiration from a popular TV game show, Power of 10[2] [14]. Similar to the show, a member of a community attempts to predict the results of a poll conducted earlier. A poll is a gathered opinion of the fellow community members about a particular belief. However, unlike the show, we do not conduct any polling apriori. Instead we are collecting polls of every member of a community within the game.

From our pilot study, we found that people found the task of predicting the exact number of fellow members who agree for the given belief. Therefore, we build our game mechanism around this task of guessing the number. Moreover, we observed that predicting exact number is a very difficult task and may reduce the fun element of the game. Therefore, to allow a certain amount of error while answering a given belief, we ask the number of fellow members in percentage.

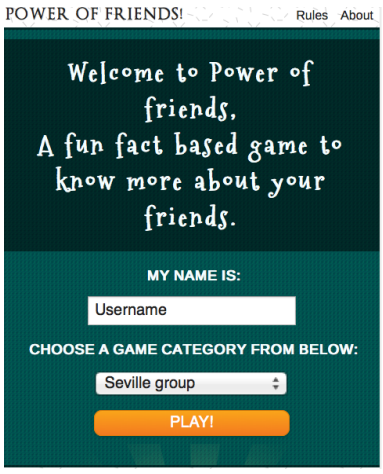

Figure 4.2 A snapshot of Power of Friends game: A player needs to register her name and selects her community before starting the game.

In our game, we ask a player to predict the opinion of other members of her community about the extracted belief. We determine truth value of the belief by collecting responses of all the members. For example, if Sumit is a member of a horror movie fans community and he likes the Vampire Diaries show,

[2]http://www.cbs.com/primetime/powerof10/free/

then we can ask individual members of his community to guess the opinion of other fellow members about Sumit's liking towards Vampire Diaries. We present each belief as a question to each member to provide her answer. If the player succeeds in giving an approximate guess, she scores points.

The presented Power of Friends game incorporates the following distinct features.

Engaging Gameplay

A simple strategy for designing the game could have been to ask a player what she thinks about a particular community related belief. Answering such a question is an easy and predictable task for a player making the game boring and mundane to play. Therefore, instead of asking her opinion, we try to relate the player to her community and invoke her interest about the given belief. We introduce randomness by asking a player to predict opinion on her fellow members. Randomness makes the task more interesting and challenging to accomplish, since the player uses her knowledge and understanding about others to predict the correct answer. Therefore, there is always an element of surprise in the proposed game design. As a result, the proposed game becomes more challenging and entertaining to play achieving our goal of engaging gameplay.

Note that the player may not always be able to make an accurate guess and in certain cases she may fail badly. But we believe that losing is an essential part of an immersive gameplay, which prompts the users to revisit the game again and again until she succeeds.

A gameplay that elicits truthful responses

The interesting feature of the game design is that we never conduct any poll apriori and all the interactions and voting happen within the game. We do not disclose the count and identity of the players who had already played the game. Therefore, a rational response of a player to achieve high scores in the game is to play honestly and respond with correct judgments about the given poll. Moreover, there are some beliefs which if asked directly would not reveal true values. For example, if a person is asked about his drug habits, very high chances are that he will deny any intake of drugs. But if the same question is asked indirectly involving the community members that how many people in the community take drugs, the chances are high that the truth about the belief will be revealed. We believe our approach has potential to uncover such beliefs within the community by keeping identity of the players anonymous. In addition, since our approach is not asking to predict opinion about one particular member, individuals find themselves comfortable to reveal their actual thinking.

We classify every belief into three categories of false, neutral and true. These categories define how the belief is considered by the community members. The belief is considered true if it is prevalent among a majority of the members, irrespective of its actual truth-value. Note that in certain cases, there might be a difference between the actual truth-value and the value accepted by the community for a given belief. Even then we consider it as an essential feedback which reflects the differences in the community opinions. For example, a very famous Indian folklore is, *"It rains when peacock dances"*, when the belief may not be true at all. Majority opinion will decide the category of the belief.

When a player gives her answer for the given belief, we check the category of the belief in which it falls. After collecting all the responses of players, we assign that particular category to the given belief which has received maximum number of votes. Taking majority opinion removes polls where players are not sure of their answer and tend to give a neutral or wrong response, thus achieving our goal of obtaining truthful responses. We will explain more about processing majority opinion in the later subsection 4.4.2

A gameplay that is asynchronous and social

We employ an asynchronous approach where only one player is completing the given task against the previous game challenges and players opinions. The advantage of our single player game is that a player need not wait for other players to join her for the game. Moreover, the approach helps the player utilize her social network to guess opinion of the closed community to which she is associated and knows it well.

4.4.1 Gameplay

The game play is simple. Each game consists of 10 random beliefs related to either the interests and preferences of the community members or to the community itself. The facts related to a community are presented as questions to its members in the game. Steps for playing the game are given below:

1. First a player goes to the game website and selects her community.
2. Each belief related to her community is presented to the player as a game challenge sequentially.
3. The player provides her opinion about the given belief by using the sliderbased control. This type of interaction avoids spelling based or other linguistic errors that may happen with text-based interaction present in the existing games like Collabio.
4. For each vote, user gets score based on how close her opinion is with her fellow community members.
5. Player proceeds further to complete her game session.

Figure 4.3 shows the screenshots of the Power of Friends game in action. The player is asked to give her opinion on a fact related to her fellow member Sumit.

4.4.2 Validation(Post processing)

To determine the category of the given belief as false, neutral or true, we process the responses collected from the players. A player gives her response on a scale of 1 to 100 by using the slider based control. This scale is divided into three approximately equal parts to calculate the belief category. A belief is categorized as false if the given answer lies in the range of 1 to 35, neutral for the range of 36

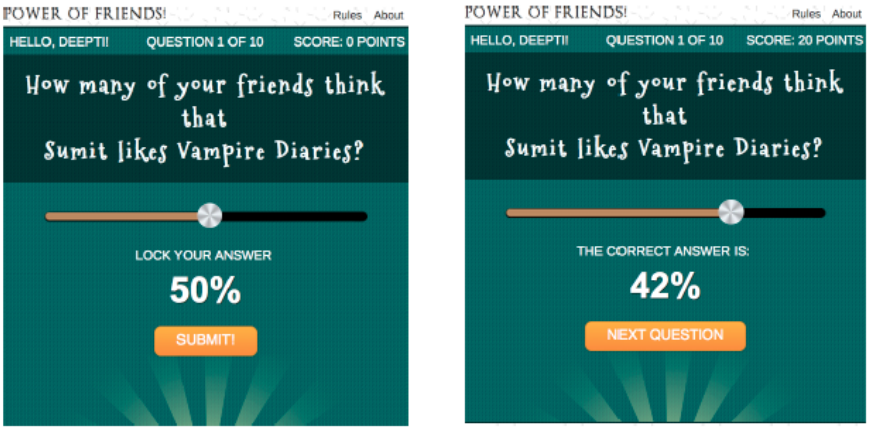

Figure 4.3 Game in action. a) Player makes a guess about her friends thinking about a given belief. b) We reveal the correct answer to the player and give scores accordingly.

to 65 and true for 66 to 100. Every response is assigned to one of the three categories based on the scale division it covers. From the collected responses, we iteratively determine the category to which majority of the responses belong. Response of every new player is compared against the majority opinion. The player is rewarded points based on how close her answer is to the community thinking. After scoring her opinion, her response gets added to the existing list of responses. We will discuss how a player is scored for her response in the next subsection.

However, problem occurs at the beginning when no player has played the game. First player should realize that someone has already played the game and her response is being compared against the previous responses. To mitigate this problem, we assign a random percentage value to each belief in the beginning and the first player of the game matches her answer against it. This random value is then discarded and does not contribute in the calculation later. The second player, thus compares her vote only against the first player and this continues for all the upcoming players where the new response is compared with the earlier collected answers of the players. Once all the players have played the game for a given belief, the category of majority opinion determines the truth value of the given fact.

4.4.3 Scoring

Scoring is based on how accurately the player guesses the majority opinion of her community. When a player gives her response for the given belief, we determine the category of her response. This category is checked against the category of the majority opinion. Majority opinion for a belief is formulated if one category obtains maximum number of votes from the community members. The player is given scores by calculating the gap between her response and the average of all the previous responses making majority opinion. The player scores maximum 100 points if she makes an accurate guess of the ma-

jority opinion (refer Figure 4.4). Moreover, the scores decrease by 10 points for every increase in 5% difference. For example, when the difference between the player's vote and the average opinion is 6%, player gets 80 points. The player scores least points(value 20) if her guess lies in the range of 40% and her answer is considered invalid if it is beyond 40%.

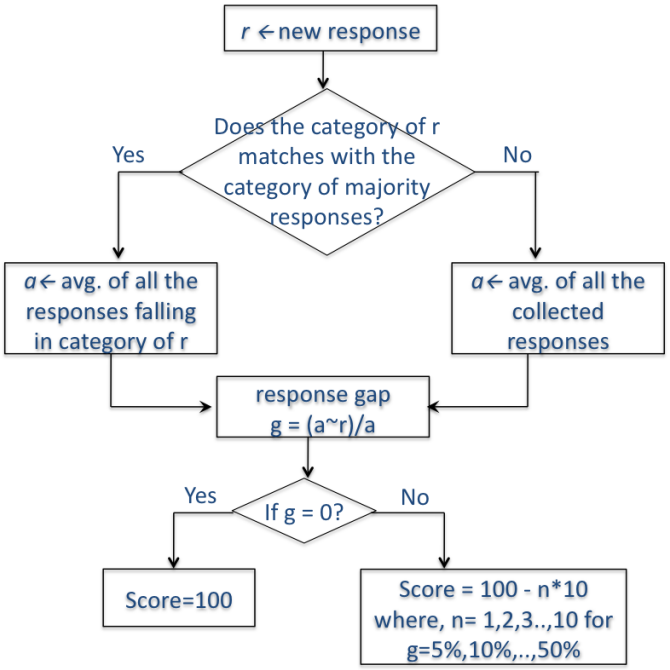

Figure 4.4 Flowchart of how scoring is done for every new response

Let us run through an example to illustrate how the processing of responses is done and how the players are scored for their response (refer Table 4.4). Consider the belief, Sumit loves Vampire Diaries. Let the first assigned random value is 48%. We will see the working of our algorithm for a few iterations with every new player.

1. Assume that the response of the first player is 64% which lies in the neutral belief category. The response of the player differs by 33% from the random value, therefore, she scores 30 points. This random value of 48% is discarded now and not used in the later iterations.

2. Now, suppose the response of second player is 20%(false category). The difference of second player with the average of collected responses is 68% and she scores zero points. Please note that the game has collected only one response so far, so the average also remains same as that of the value, that is, 64.

3. Suppose the third player gives 40%(neutral category) to the given belief. In this case, two responses so far obtained 64% and 20% do not make any majority but her vote matches with the

Previous responses (in percent)	Count of responses in each category	Response of a new player (in percent)	Average of the majority responses	Response gap (in percent)	Scores of the player
48(random)	-	64	48	(16/48)*100 = 33	30
64 (48 discarded)	f=0, n=1, t=0	20	64	(44/64)*100 = 68	0
64, 20	f=1, n=1, t=0	40	64	(24/64)*100 = 37	20
64, 20, 40	f=1, n=2, t=0	36	52	(16/52)*100 = 30	40
64, 20, 40, 36	f=1, n=3, t=0	67	46	(21/46)*100= 45	0

Table 4.4 An illustration on processing the players responses and scoring them for their responses

category of first player. Therefore, we score the player based on her response difference with 64%(average same). Players gets 20 points as the response gap is 37%.

4. Supposing fourth player gives 36%(neutral category). Now, neutral category got the maximum responses, therefore, the response gap of the player is calculated with the average of previously collected neutral responses. The average of 64% and 40% is 52%, and the response gap is 16%, thus player scores 30 points.

5. Suppose fifth player gives 67%, making the response gap of 45% with the average of three values collected in neutral category(64, 40 and 36). The player scores zero points for her answer.

This continues for all the upcoming players. Finally, when the game stops, we get the truth value for the given belief in Sumit's community, that is, whether community members of Sumit think that he likes Vampire Diaries or not. In this scenario, the fellow members of Sumit have the belief that he is not fond of Vampire Diaries but moderately likes it.

4.4.4 User study

We conducted a focussed group study to test the following proposed features of our game:

1. Does the gameplay elicit truthful responses from the players? We aim to test if accuracy of the beliefs generated with our proposed approach is good.

2. Is the gameplay engaging? We aim to test if the participants find the game more engaging while predicting crowd thinking than giving individual opinion.

Setup

A web-based prototype of Power of Friends game was developed in Javascript and HTML5. Standard desktop systems with resolution of 1024x768 were used by the participants for the study. The current

prototype has beliefs for few communities selected by us. The game contains questions regarding preferences of the selected community members. For this study, we did not extract the questions from any social networking websites. Instead, we asked few members of the community to generate questions about all the members of the community. The questions were related to likings, hobbies and daily activities of their community members. One of the selected group was of kids. We extracted all the facts related to this community by having informal conversations with them.

Participants

We selected 67 participants (40 female) and formed seven communities, C1, C2, C3, C4, C5, C6 and C7, out of them (Table 4.5). Each participant belonged to only one community. The participants were invited by sending mails and calling over phone. The average age of the participants was 22 years, while the youngest participant was 9 years old and the oldest was 28 years old at the time of study. First five communities were selected from our university campus and the last two from outside the campus. The campus communities were freshly formed and people know each other from last one or two years. But the remaining two were old communities where people know each other from a long time. Members of first five communities were staying together in the campus while that of C6 and C7 had not met from a long time but were in contact with each other. We sent the URL of our game to the members of C6 and C7 and took online feedback from them.

ID	Selected Communities	No. of members	Bonding	Community age	General topics of discussion
C1	Research students	15	Moderate	2	Science and Research related
C2	Masters incoming batch	7	Strong	1	Course related assignments, personal queries, fun activities
C3	Professionals	8	Low	2	Technology review, time and stress handling
C4	Masters 2010 batch	7	Strong	2	Job interviews, subject study, personal queries
C5	Kids	9	Moderate	8	Homework, games, jokes, fun activities
C6	B.Tech. 2006 batch	13	Strong	6	Job interviews, personal queries, fun activities
C7	School friends	12	Moderate	10	Jokes, fun activities

Table 4.5 Description of the selected communities

We have specifically chosen seven different communities such that there exists variations in terms of age and bonding among the participants. From the pilot study, we obtained an important finding that the accuracy of a player depends on her level of bonding with other community members. Therefore, in this study we were interested to see the accuracy level of each community based on the closeness among its

members. To know the interaction level of members within the community, we asked different questions from them while collecting the demographic information. We inquired about their bonding level with other fellow members, the topics of discussion in which they are involved and the duration from which they know each other. After collecting the response of every member of a community, we define a bonding value to each community by taking the majority (Table 4.5).

Procedure

The study was conducted in two phases similar to the procedure followed in pilot study (section 4.3.1). Participants were given short demo on how to play the game and how to give their response using the slider bar control. Participants were explained the rules associated with the game. Each player is asked to play a game related to her community. Each game consisted of 10 questions containing random beliefs related to the members of a community. Figure 4.5 gives a description of the variety of questions asked from the players inclusive of all the communities. Each game session lasted for an average of four minutes. At the end, we collected qualitative feedback from the participants about the game through open interviews. Participants were also encouraged to think aloud while playing the game. This was done to collect their experience about the game play and to know whether the players are liking the game features and are able to approach the questions or not. We present the cumulated information of user experience in the later section on user satisfaction.

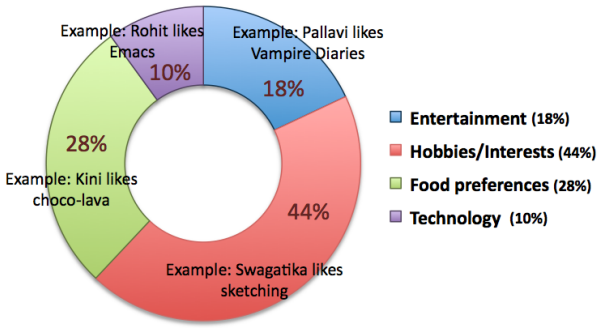

Figure 4.5 Division of different types of questions covered in our game. An example is given for each topic.

4.4.5 Results

We report our findings based on the data collected from 67 game sessions in terms of the following:

1. *Accuracy* to accomplish our first objective. This is determined by the total number of correctly identified beliefs by each member of the community.

2. *User satisfaction* to accompish our second objective. This is determined by the responses of participants collected during the questionnaire session.

Accuracy

Each player's response is checked against the category receiving majority of opinions. If it matches, her opinion matches with the majority, and hence we consider that she has identified the correct value for the given belief. Performance of each individual of a community is determined by calculating total number of responses that match with the majority opinion. Moreover, to determine the performance of whole community, we calculate the average of performance of each individual member of the community. Once all the participants of a community have played the game, category of each belief is determined by the category receiving majority votes.

We analyzed the performance of each community as shown in Table 4.6. It describes the average number of beliefs correctly identified by each participant of the selected communities. Figure 4.6 presents the same data in form of a bar-graph, where each bar represents the average of correctly identified and wrongly identified beliefs by each community. The figure also demonstrates that 88% of the participants (59 out of 67) correctly identified 50% of the facts (5 out of 10).

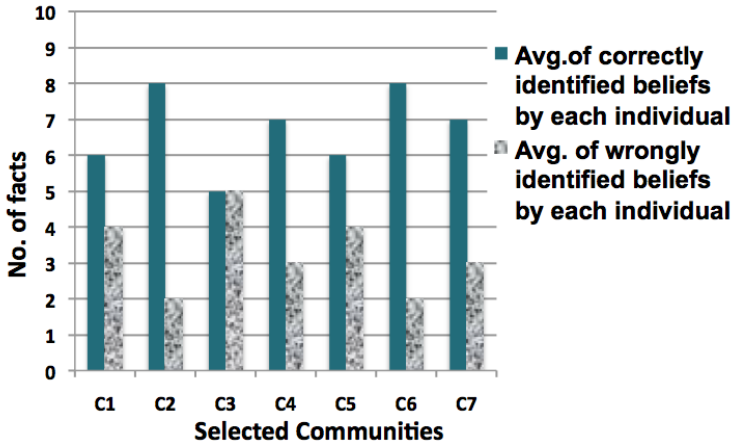

Figure 4.6 Accuracy of each community: 88% of the participants correctly identified 5 facts

As we can see from Table 4.6, the responses of the communities C2 and C4 were very accurate, while the accuracy was lowest in C3. The reason behind such a performance by C3 members was that the members never really interacted much with their fellow members, whereas the members of C2 and C4 are very closely connected with each other. This result reflects that the accuracy of the validation tasks depends upon the involvement of a person in the community activities. If the member is active within the community, she knows more about her fellow members interests, and therefore, can provide more information about the same.

From the table, we found that the accuracy of each community differs. We plot a bar-chart, Figure 4.7, to show the correlation of accuracy of the community to identify beliefs with the bonding level

Community ID	Average accuracy of each community member
C1	6/10
C2	8/10
C3	5/10
C4	7/10
C5	6/10
C6	8/10
C7	7/10

Table 4.6 Community C2, C4, C6 and C7 were more accurate in identifying correct beliefs.

within the community members. The bar is plotted by taking average of performance of all the communities where members are connected with each other with same bonding type.

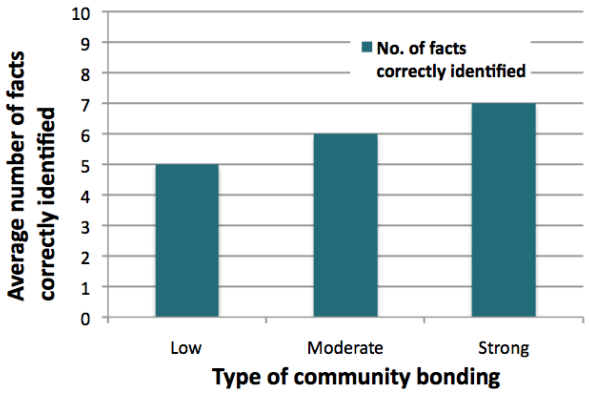

Figure 4.7 The chart shows that the accuracy of community to identify beliefs increases with the level of bonding within the community members.

User Satisfaction

We conduced interviews at the end of the user study to get qualitative feedback about the game. The first question asked was to describe the experience of playing the game. Eighty percent of the users found the game simple yet challenging and they wanted to play the game regularly. The most enthusiastic players of the game were kids while professionals find it most tedious. One member of C3 community said, 'It requires a lot of thinking. I wish I knew my coworkers better'. Another participant from C1 stated, 'It's kind of fun to see how accurately my thinking aligns with my friends'. Sixty percent of the participants also responded that the game has helped them to know new facts about their friends.

On asking about the likings for the game features, 84% of the users responded with the guesswork of thinking about others as their favorite feature; 92% of the participants liked the simplicity of the interface design; while 82% participants appreciated the selection of the questions. Finally, we asked their suggestions on further improvements and additions in the game design, to which 40% of the participants wanted to extend the game to a multiplayer setting so that they can compete with their friends and have more fun. Ten percent of the participants wanted to introduce time-based challenge. Few users also suggested including more personal or intimate questions in the game, to make it more exciting.

4.4.6 Discussion

We observed that it is difficult for the participants to remember the game mechanism of thinking about others while playing: about 25% of the participants got confused while playing and thus needed help to remind them the game strategy (of predicting the opinions of other community members).

We also asked individual opinion of the participants while providing the community opinion for the given fact. The members of the community C3 hardly know each other's preferences in the community. Therefore, their individual thinking was same as that of the group thinking for 90% of the given facts (9 out of 10). While for rest of the communities, individual thinking matches with the group thinking for 70% of the given facts (7 out of 10). For example, for the above mentioned fact, *'Sumit likes Vampire Diaries'*, very close friends of Sumit in the community know the actual truth about the fact. But they also know that the truth prevalent in the community is different. In such a case, their individual thinking differs from the group thinking.

The user study of the game design showed that the participants find the game design engaging and the game play also brings out truthful responses for the given facts from them. Since this was the first exploratory study of the presented game design, we considered a convenient sample and did not conduct a field study. The results of the study motivate us to conduct another study with an objective of testing the differences in the opinions of different communities for the same set of facts.

4.4.7 Limitations of the study

There are certain limitations of the study we conducted to test our approach, Crowd Consensus, and our game, Power of Friends. Firstly, we considered only those communities where members are directly connected to each other, and thus are aware of each other's preferences. However, such direct mode of interaction is not possible for bigger communities, say, Telugu speaking people, as bigger communities follow indirect mode of interaction where people are not directly connected with every other member but still share common beliefs among themselves. We are currently working on the extension of Power of Friends, where we will test our game with larger community sharing particular interests.

Moreover, the study was conducted on an online game prototype with a set of questions manually collected from Facebook. We are planning to conduct our next study on Facebook with the existing

communities such as a community of CHI-2014 attendees. This natural setting will provide us better understanding of the feasibility of our game.

4.4.8 Conclusion and Future Work

We demonstrated a simple but entertaining game design, Power of Friends, to extract the community knowledge from its members. The game is based on our proposed approach Crowd consensus where each member is required to predict the opinion of her fellow community members about a given belief. We describe the complete design of the Power of Friends game and report the results of the empirical studies. Empirical studies of Power of Friends have been extremely positive. More than 80% of the users found the scoring and user interaction elements of the game highly enjoying. More than 80% participants correctly identify more than 50% of the beliefs.

Although we modeled the presented approach in the form of a game, we strongly believe and envision that our approach can be applied to other domains, not just friends getting to know better and having fun, but also professionals reaching an agreement or groups converging on a decision or to other human computational systems such as Amazon Mechanical Turk. Moreover, our approach may also provide a solution to reduce the number of iterations required to complete a crowdsourcing task. The reason being that even a lesser number of people forming a community should be able to predict the opinion of their fellow members accurately.

This page is intentionally left blank.

Problem 2- Extraction of named entity relationships

This page is intentionally left blank.

Chapter 5

uPick: Crowdsourcing Based Approach to Extract and Validate Objective Facts

5.1 Introduction

Despite the advancement in the information extraction area, the task of identifying named entities relations within a text document remains a significant challenge. In Chapter 2, we have discussed various automated techniques to extract named entities and their relations for different domains like medical and newswire[37, 46, 56, 67]. In the chapter, we have also mentioned various problems associated with the existing techniques because of which a system fails to achieve high accuracy. The most prominent among them are dependency of the system on external resources like Wikipedia, and the poor scalability of the system on different domains.

Existing automated approaches lack human precision and they also struggle to handle erroneous documents. The situation could be improved if we utilize human judgments on the extracted relations among named entities. However, humans find the task of filtering (analyzing the accuracy) cumbersome and not particularly engaging [109]. In Chapter 3, we have discussed games that are developed to extract relations among named entities, e.g., game of protein folding OntoPronto [89]. We have also discussed various challenges to design human computational games that make mantaining the trade-off between the fun element and the accuracy of the output data difficult. Therefore, we aim to design an immersive environment that motivates a human to particularly engage herself in the tasks of filteration of named entity relations.

In this chapter, we propose a crowdsourcing-based approach to improve the accuracy of the generated relations from the existing extraction techniques. Our idea is to gather judgments on the extracted relations of an article from the interested users. By contributing, the users in return remember the facts related to a document. This chapter presents the complete design of the approach along with a user study done with twelve participants. Results show that the users rated the proposed system positively and were willing to contribute their time and energy for the task. Moreover, the accuracy of the relations collected by using our approach is more than 75%.

The rest of the chapter is organized as follows: Section 5.2 describes the architecture of our proposed scheme, uPick. The experimental design set up to test uPick approach is explained in Section 5.3 and the results are presented in Section 5.4. The paper is concluded in Section 5.6, stating all the possible future work of the proposed design.

5.2 Our approach: uPick

We propose a simple single-player human computational game called *uPick*, to extract objective facts from documents. The approach divides the task of generating named entity relations into two phases: extraction of relations from the given text, followed by filteration and verification of the generated results. To minimize the cognitive effort required for a user to play the game, our approach engages users only in the verification of the relation whereas, the entity relations are extracted from the given text by using some existing automated techniques. The proposed approach involves only the interested users reading an article towards filtering the named entity relations retrieved from it.

The system architecture is explained in Figure 5.1. We first extract all the possible relations from the document using Part-of-Speech (POS) Tagger [42] and some rules based on English language construct. Please note that we have implmented a fundamental automated technique of POS tagging in our system. We did not complicate our system with other machine learning or Natural language processing approaches to produce accurate results. Instead, the gaming approach in our system accounts for the accuracy of the generated relations. Moreover, system interference is not required once all the relations are generated.

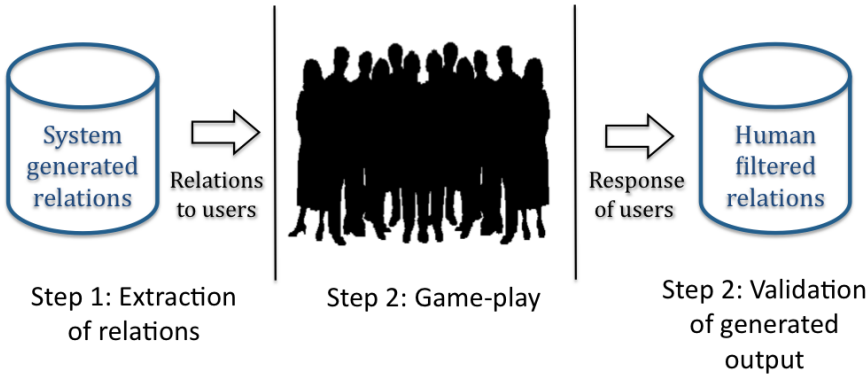

Figure 5.1 System architecture of uPick presenting the extracted relations to the users as a challenge

The extracted relations generated using automated technique are then presented to every interested user in the form of a challenge for filtering. User then identifies all the valid relations among the presented ones and ticks them. We are willing to take the consensus of each reader for each relation

derived from the document using the proposed gaming approach. To encourage active participation from the users, we utilize a reputation-based system where contributions of individual users are listed publically for others to view in terms of scores. We believe that doing so will initiate competitive nature among the players(interested users) and will motivate them to contribute heavily towards the given task. Finally, we verify the collected responses by cross checking them with the responses of other players. We filter out all those relations of a document, which are voted by majority of the interested users.

To elaborate the architecture, uPick involves the following steps: 1) Extraction of relations 2) Game Play 3) Validation, and 4) Scoring. Let us explain them one by one.

5.2.1 Extraction of relations

We first extract all possible relations among named entities from a document using any of the existing basic automated technique. In the current prototype, we have used the Stanford Part-of-Speech tagger [16] to determine and tag the basic constructs of English sentences such as noun, verb, adjective, adverb, etc. To extract relations, we utilized eleven relation extraction rules proposed by Chen et al. [27] along with some other heuristics based rules. These rules help to identify a named entity, its relationship type and the corresponding attributes from the tagged constructs. Named entity relations are then generated automatically for a given text. Figure 5.2 shows the working of the automated techniques used to extract the relations automatically.

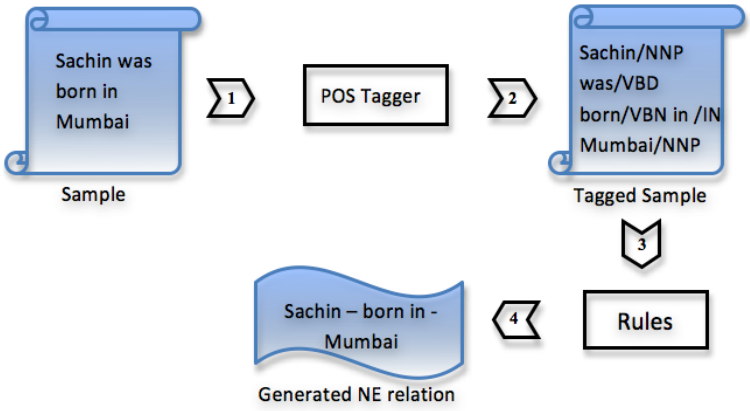

Figure 5.2 Process of extracting named entity relations on one sentence using automated techniques

Please note that we have made use of corpus independent techniques, namely, POS tagging and rules based on English language structure, which by themselves, provide very less accuracy. We believe that our system along with human iterations will improve the accuracy of automatically generated relations over time.

5.2.2 Gameplay

The proposed game is a single player game that challenges users to provide their knowledge of a recently read online document. The challenge is in the form of a set of questions that the users are to validate. The facts related to the document are presented on a browser interface with the document to be read followed by the corresponding named entity relations. The users provide their judgments about the authenticity of each system-generated relation. The relations that are not selected by the user are considered to be irrelevant and considered invalid. Following are the steps involved to play the uPick game:

1. First a user goes to the uPick website and reads a given article.

2. All the relations extracted from the document are presented to the user as a challenge at the bottom of each article. Participation is not compulsory and the user can participate provided she is interested to play.

3. The user must tick all the facts that she thinks are true for the given article. In the current prototype, we present all the extracted facts with check boxes, where the user responds with her judgments.

4. For each judgment, user gets some score based on a majority voting explained in the later subsection.

Figure 5.3 shows a snapshot of the uPick game in action where a user is asked to respond with correct facts about Sachin Tendulkar.

In future, we are planning to design better interactive features, such as providing three filters to ask user to vote from: True, False, and Do not know. We would also like to add a provision at the end of the game play where, the player can compare his performance with his friends and check how well he has performed with respect to them.

5.2.3 Validation (post processing)

For each document, we collect valid relations from automatically generated relations (generated in extraction step) with the help of human experts. These valid relations are stored in a database. Once the game has been played by a significant number of players, we compare the collected responses from each game against the expert corrected facts stored in the database and filter out erroneous response data. The relation instances having a majority of votes are taken as true facts corresponding to the document. Therefore, retrieving such facts produces filtered relations associated with the named entities appearing in the document, which are then stored as valid relations in the database.

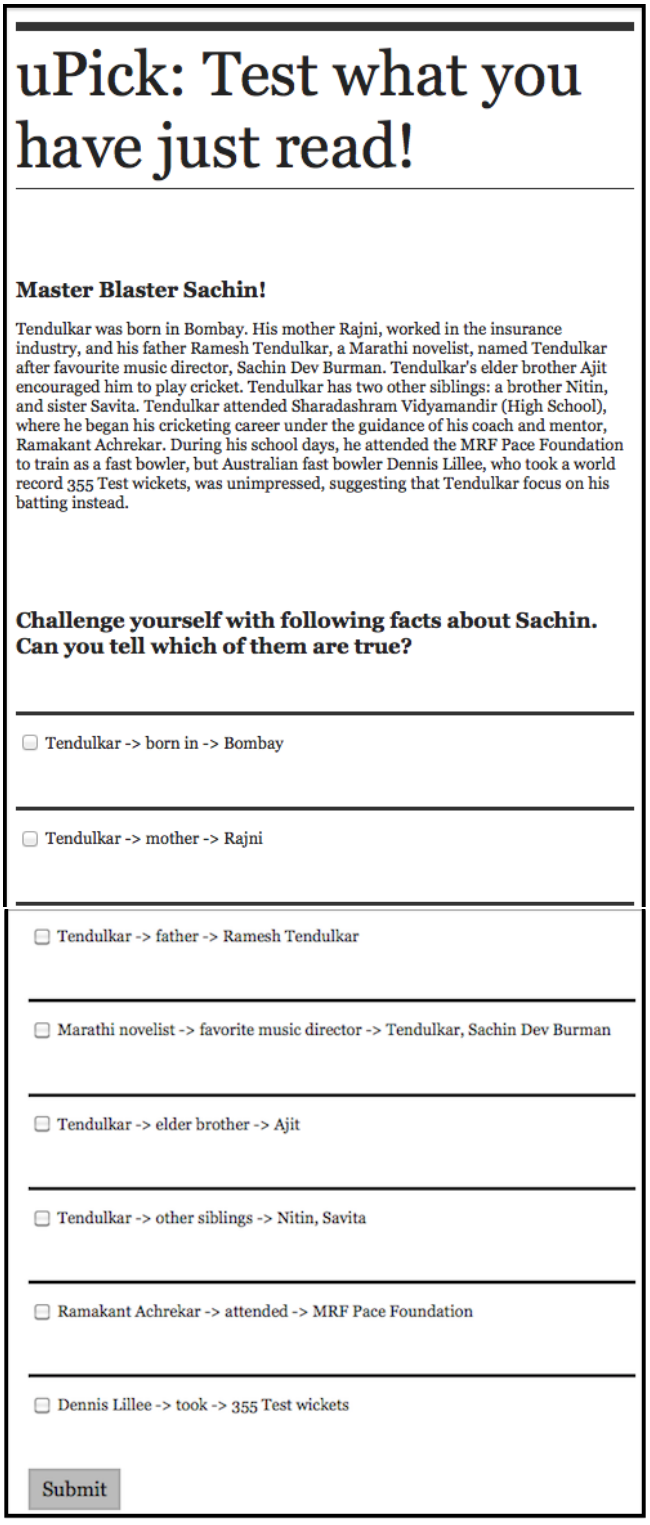

Figure 5.3 uPick game in action: The player is challenged with a set of questions related to Sachin Tendulkar

5.2.4 Scoring

Our system is based on obtaining majority voting from the users to filter out relations present in the given document. A score is awarded for every user response (i.e., filteration of a single fact from the presented list) if her response matches with the majority. For example, if majority of the users who have played the game vote for a particular fact, and the current user response matches with majority, then score is awarded. Majority in our game is more than 50%. However, for the first user, her responses are compared against the expert corrected set of relations.

5.3 User Study

A web-based prototype of uPick game was developed in Javascript and HTML5. Standard desktop systems with resolution of 1024x768 were used by the participants for the study. To perform user testing we selected four articles on Ashok Maurya, Sachin Tendulkar, Shahrukh Khan, and Sonia Gandhi from Wikipedia and named them as D1, D2, D3, and D4 respectively. All the named entity relations are extracted for these selected documents by using the technique discussed in Section 5.2.1. Table 5.1 shows the total number of extracted relations for each document along with their accuracy, verified manually with the help of an expert.

Accuracy of automated technique used	**D1**	**D2**	**D3**	**D4**
Total number of presented relations	37	39	40	33
Valid relations from the extracted facts	24	24	23	16
Invalid relations from the extracted facts	13	15	17	17
Accuracy (Valid relations / total relations)	65%	61%	57%	49%

Table 5.1 Accuracy of the extracted relations from the selected documents set using automated system

Each document gives different number of relations, out of which some are valid and the remaining are invalid. Valid relations are those which are complete, that is, contain a subject, a relation and an object and convey a correct meaning. We invalidate the relations, if they are either incomplete (subject or relation or object missing), or do not convey any meaning. Therefore, the accuracy of the generated relations is the ratio of valid relations to the system-generated relations.

The extracted relations from each of the four documents are then used to formulate a challenge (one per document) where we ask the user to verify the authenticity of each relation after she has finished reading the document, that is, whether the relation holds true or false for the given document.

5.3.1 Study Objectives

We conducted a supervised laboratory study of uPick with the following objectives:

1. Does the uPick scheme generate more accurate relations than the automated scheme?

2. Does the uPick system provide an interesting environment to the users such that they are motivated to contribute to the task frequently?

5.3.2 Participants

We conducted a supervised laboratory study to test the accuracy of our uPick scheme against an automated system. We recruited 12 participants from our university campus by sending invitation emails. The average age of the participants was 15 years, the youngest participant being 10 and the oldest being 29. Four of them were male and eight were female. To evaluate the efficiency of our proposed approach, we required a sample population who read online. Therefore, we targeted a specific population group of younger people to perform our experiments.

5.3.3 Procedure

Our usability test consisted of two sessions that span for an hour. The first session was dedicated to registration and training. First, the participants got an introduction to the study, and the procedure to play the game was explained later. The second session was dedicated to the actual gameplay. Each participant was given a task of reading two of the four test documents (randomly picked) and then to solve the accompanying challenge. Pseudo-randomization of document selection enabled counterbalancing and helped to minimize the learning effect. Therefore, six different participants in our experiment evaluated each document. At the end of the study, participants were asked to fill a questionnaire for the qualitative analysis of the proposed system. The questionnaire responses were followed by a small interview with each participant.

5.4 Results and Discussion

We report our findings in terms of the following:

1. Accuracy (defined by total number of relations identified correctly)

2. User satisfaction (by users feedback)

5.4.1 Accuracy

For our uPick scheme, we measure the accuracy in terms of the total number of relations correctly identified by the participants for each document when compared against the expert opinion. Table 5.2 shows the performance of the participants and the accuracy of the system achieved by filtering the relations with majority votes.

Accuracy of uPick Scheme	D1	D2	D3	D4
Total number of presented relations	37	39	40	33
Correctly identified valid relations	19	18	19	15
Correctly identified invalid relations	12	12	16	15
Incorrectly identified valid relations as invalid	5	6	4	1
Incorrectly identified invalid relations as valid	1	3	1	2
Accuracy (Correctly identified relations/total relations)	84%	77%	87%	91%

Table 5.2 Accuracy of uPick scheme considering majority votes of the participants

As discussed, for a given challenge users gave their judgments by marking the correct relations and leaving the incorrect relations unmarked. However, it is possible that a user marks an invalid relation as valid or leaves a valid relation unmarked considering it invalid. In both possibilities a user is not able to identify the given relation correctly. Therefore, we analyze the accuracy of the uPick system by the total number of correctly filtered facts, both valid and invalid, by calculating the majority votes of the participants.

From the table, we can observe that the responses of the participants were fairly accurate for all the four documents. This gives us an insight that the users are able to perform the task of filtering in a significantly efficient manner. Therefore, we can say that the uPick scheme involving human intervention improves the accuracy of an automated system.

5.4.2 User Satisfaction

At the end of the user study, we collected oral feedback from every participant about the presented scheme. Seventy-five percent users(9 out of 12) found the system simple and easy to use and they were willing to contribute their time and energy for such tasks, provided the presented documents are of their personal interests.

In our scheme, the user has full freedom to perform the task in a manner, which please them. We observed that, forty-one percent users(5 out of 12) performed the challenge after reading the complete document; thirty-three percent users(4 out of 12) preferred to read each paragraph and then performed the challenge; and twenty-five percent users(3 out of 12) located the sentences of the given document based on the presented relations and then validate the relations. When asked how such a system can be helpful to them, ten participants replied that it would help them to remember facts related to the concerned document. Four users suggested that such a scheme would be helpful to avid users to verify and extend their knowledge in an entertaining way.

Two users didn't appreciate the presentation of the challenge at the end of the document; instead they wanted a flexible scheme wherein the challenge related to the document is presented after a random paragraph containing the facts of the previously read paragraphs. According to them, such randomness will stimulate the task even more in terms of finding challenge related to each paragraph and will yield

more learning environment. However, three participants didn't find the present game design particularly engaging and suggested a few alternative designs as puzzles and object finding games.

With the performed study, we found that the proposed scheme is not able to entice users as a game because of lack of fun and enjoyment factors. However, it is very useful for creating a learning experience to the users in an interesting way.

5.5 Limitations of the study

Our study has certain limitations, which we mention in this section. First, we tested our approach only on four short documents having 5 paragraphs at max. with 12 participants. Additionally, the documents were manually created by taking text from the corresponding Wikipedia pages. Given these limitations, we can not provide proper justification to our research objectives. A more extensive study with a large group and bigger documents in a natural work space of users is required to substantiate our research objectives with this approach. Moreover, in the current prototype, users are provided scores based on their performance for the given challenge, but we will facilitate our scheme with leaderboards in the near future to invite more frequent user contributions.

5.6 Conclusion and Future Work

In this chapter, we proposed a crowdsourcing-based scheme to extract the objective facts about named entities from a given document. We developed a system called uPick, which extracts all the relations from a document using the fundamental automated approach of tagging the sentences and then applying few rules. Our idea is to gather the judgments on the extracted relations of an article (system generated) from the interested readers, and thereby filter out the valid relations from them. Below we mention two essential benefits of the scheme.

Effectiveness: The game is designed in such a way that it requires minimum human cognitive effort and time. The users need to give their responses by clicking on the fact they find related to the document. We believe clicking the options rather than writing makes the task easy and interesting. Moreover, the approach does not depend on any external resources, and therefore, can easily scale well for any domain and corpus.

Generalization: The proposed approach is independent of the document language and to any data corpus. Such a generalization will require changes only in the rules to extract relations from the document because the rules are dependent on the grammatical structure of the language of the given document. We believe our proposed approach can be implemented to validate information of different types, for example, validation of anaphoras and sentiments related to a document.

The user study showed that our approach provides an interesting environment to the users to read documents, however, it fails to provide the required fun element related to the associated task. Amongst the factors to persuade human participation, we observed that making a task interesting is the most important factor. Therefore, as future work, we plan to extend our system to an interesting game for performing the task of filtering Named Entity relations in a Crowdsourcing environment.

With the feedback we have obtained during the study, the proposed approach could be useful for different applications such as: to develop a Question-Answering system on individual documents where the relations collected by our approach could be used to answer user queries. Moreover, as users found the approach useful to learn facts related to a document, an interactive learning system could be developed using our proposed approach. This application could be useful in a remote classroom environment, where the performance of a user needs to be evaluated and presented as her progress report card.

Problem 3- Re-narration of the web documents

This page is intentionally left blank.

Chapter 6

Reviewing Accessibility Issues With the Web

6.1 Introduction

Penetration of the web has crossed geographical boundaries. It has connected people from different parts of the world, thus enabling them to share their views and discuss topics of various concerns through web applications. Since the web allows freedom of speech (in the form of publishing content), there exists diversity in the presentation of the content. The web has become a wide reservoir of information covering all the possible domains, such as, medical, sports, and geography information etc. The presence of the web has witnessed a shift in performing our normal activities from accessing traditional media to accessing web applications having highly interactive and massive multimedia content. For example, online shopping services have become very popular into fulfilling our daily requirements. However, despite the phenomenal growth of Internet, major population groups of the world remain outside its influence. One of the goals of Web 3.0 is the Web penetration to our daily activities, which is still only 32%[9]. Poor accessibility of current Web resources significantly contributes to this problem. A page is considered accessible to a user (reader) if she is able to truly access and understand the conveyed information.

There are various factors which contribute to the variety in expressions on web-pages and hence lead to inaccessibility of the web [4, 49]. One of the reasons has been the social and cultural differences across regions. For example, the food habits, dressing style and way of living may differ in different culture. The other reason is the difference in the geographical conditions adding to contextual differences in expressing the information. For example, a web page, written in English, illustrating home remedies for medical diseases, mentions intake of coconut water for body weakness in summers. This information is not much useful for people belonging to the regions where coconuts are not produced. For them a possible remedy could be the intake of lemon water. Here, even if the page is written in English and the user understands English, the unlocalized information contributes to contextual difference of the data.

Another major factor causing inaccessibility of the web content is the language barrier: an individual is either illiterate and thus can not read or she finds herself illiterate with respect to another

language. Most of the web users are facing the second type of language inaccessibility which we term as *"non-literacy"*. The reason being that the overwhelming fraction of the Web content today is in those languages that are inaccessible to a large population of the world. For example, as of December 2011, no Indian or native African vernacular contributes even 0.1% of the total number of Web pages (Figure 6.1)[6]. Moreover, according to the recent consensus, only about 10% of India's approximately 880 million literate people speak English [7, 11]. This implies that majority of the Web content is inaccessible for nearly 800 million literate people in India. In addition to this, physical disabilities such as: visual or auditory impairments, poor cognitive skills, further add to the web inaccessibility.

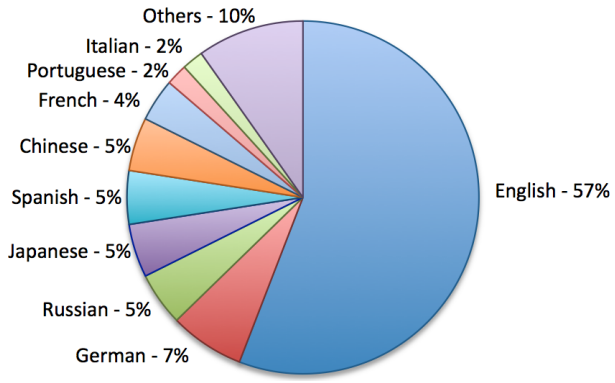

Figure 6.1 Distribution of the web pages in different languages.

Power of the web lies in its potential to provide universal access to everyone regardless of any barriers such as limited Internet connectivity, physical impairment, linguistic differences, and social, cultural and geographical factors. To realize this potential, the web should be designed to make it usable to all groups facing different accessibility limitations. Such a web design to unite excluded groups is referred to as *Designing for Social Inclusion*[15]. To achieve social inclusion, the web-page authors and developers have to follow the Web content accessibility Guidelines (WCAG). The existing approaches are either based on checking whether the W3C guidelines are followed properly by the web-pages or using certain software/browser plugins to convert the page into an accessible form. Significant work to tackle the issues related to accessibility include: video with text files (srt), text-to-speech, and speech-to-text conversion, and some elementary changes in the web page elements(e.g., font size, page color) at client end.

Most of the efforts in making the web accessible by everyone is in the direction of improving accessibility for physically impaired population. Several assistive technologies such as screen reader and voice recognition system[24, 55, 77, 105] have been developed to aid such people. Traditional approaches to solve lingual accessibility problem consist of automatic translation services for cross-lingual web resources such as online translation services using Google translator. But non-literacy has not gained much attention so far. More recently, social collaboration based approaches have emerged as an inter-

esting and popular alternative to create content on Web using humans. For example, social collaboration sites like Wikipedia, blogs etc. utilize the wisdom of crowd for generating content[94]. Under the rubric of social accessibility, these approaches rely on the power of individual users to improve web page accessibility in a decentralised manner. However, there is another aspect of social accessibility which is concerned with the issue of the socio-cultural background and the geographical conditions of the content consumer; "social" in this context refers largely to the process in which content is created and managed by humans.

Each of the approaches mentioned above are important but the existing technologies do not adequately address the non-literacy and social accessibility issues. The reason being their ignorance about the specific socio-cultural and localized aspect of the reader. This also means that the contextual differences of the readers are also ignored. In the chapter, we discuss and illustrate the potential of re-narration approach to solve accessibility issues of the web. The re-narration approach allows people to rewrite different elements of a web-page to any medium such as text, video, audio or image. In the thesis, we have worked on a collaborative approach to solve the web inaccessibility. The main focus of our research is to solve the problem of non-literacy which uniformly tackles the inaccessibility caused by other factors. Therefore, we limit our study to crowdsourcing based approaches developed so far to solve the web inaccessibility issues.

In this chapter, we first explain the existing technologies developed in order to solve the problem of web inaccessibility with assistive tools (Section 6.2) and user profiles (Section 6.2.1). In Section 6.3, we propose our approach of re-narrating web using social collaboration and illustrate the approach with several inaccessibility issues. Lastly, we conclude the chapter in Section 6.4.

6.2 How is the web inaccessibility tackled so far?

The most prominent initiative towards achieving social inclusion was taken by W3C's Web Accessibility Initiative (WAI)[63, 103]. It has a model consisting of three sets of guidelines, which are described as follows:

- Web Content Accessibility Guidelines(WCAG) for web page authors.
- User Agent Accessibility Guidelines(UAAG) for browsing and accessing technologies.
- Authoring Tools Accessibility Guidelines(ATAG) for tools to support web content creation.

However, the WAI model has certain shortcomings and therefore, does not fulfill the requirements of Web 3.0[58, 59]. One of the biggest issue is that the model requires conformance of each of the three sets of guidelines, making them dependent on each other. But this dependency is very hard to follow in real world applications. For example, web authors can control the quality of web content by following WCAG guidelines but they generally have no control over the browser technologies used by the client

to access web information. Also, a web-page author may not be able to visualize the possible set of readers having different capabilities beforehand.

Earlier, the guidelines were based on the principle "one Web content for everyone", which is not considered as much effective as "the best Web content for each one"[77]. In order to meet the new principle, there have been efforts enabling people to access web by considering all inaccessibility issues individually. A new technology has developed for every other kind of disability. For example, a visually impaired person is powered with technologies like a screen reader for text annotation and accessing video media with flexible navigation[24, 55, 86]; a deaf person is enabled with speech-to-text assistive tools[31]. Some other work includes client-side browser settings with plugins, where the plugins restructure the web page elements(e.g., font size, page color) according to the user needs[77]. Some other attempts include facilitating low-literacy readers by providing them with simpler texts along with other relevant written material. For example, alleviating a reader by providing synonyms for difficult words and tagging named entities for clear understanding of the document[104, 105].

A few efforts are also made to solve the problem of functional illiteracy for documents with different social, cultural, and geographical contexts. Topac et al. proposed a framework to solve the issues with context specific information for medical domain. The approach empowers patients to understand medical specific terms in a lay person's language[96]. Borodin et al. designed a system to make the content accessible in different languages using Google Translator[21].

There are several examples of community sourced initiatives for subtitling and other needs creating meta-data of the page[78, 95]. Moreover, there are plenty of online services allowing people to provide annotations, reviews or comments about research papers, books or products. Some of them include Citability[1] to annotate government documents, Digress.it[2] and Crocodoc[3] to annotate a document, iCorrect[4] to rectify rumors and misinformation about celebrities, FinalsClub[5] to access annotations of academic courses offered by good universities. Most of these initiatives are based on the intrinsic motivation, which we have discussed in Section 3.2 of Chapter 3. Some of these applications also provide monetary incentives to perform the task, for example, iCorrect pays its contributors.

The approach proposed by Takagi et al. works within the framework of WAI guidelines by using community based collaboration[94]. Given a page without appropriate accessibility tags(alt text for images, headings), the approach uses social collaboration to generate a modified page by adding metadata to its sub-sections. The notable part of this work is the use of social collaboration where a group of supporters fill in missing metadata whenever a report of inaccessibility comes in, bypassing the page author completely. The modified page is stored on a centralised server which makes it accessible to anyone visiting the original page. All these services – reporting inaccessibility, generation of metadata

[1]Available at http://citability.org
[2]Founded by Eddie Tejada in 2011. Available at http://digress.it/
[3]Founded by Ryan Damico in 2010. Available at http://crocodoc.com/
[4]Founded by David Tang in 2011. Available at https://www.icorrect.com/
[5]Founded by Andrew Magliozzi in 2007. Available at http://finalsclub.org/

by supporters, as well as identifying existing re-narrations – are provided by a set of client-side tools that interact with the server via a set of APIs.

However, all the above mentioned approaches require understanding of user preferences to restructure a web-page and its attributes at client side. Several approaches are proposed to capture the user preferences using user profiles, which we discuss in the next section.

6.2.1 Building user profiles as an effort to solve web inaccessibility

Recently, there have been efforts to solve the web inaccessibility issues automatically with personalized user profiles. These user profiles defines the choices and capabilities of the user. The profile helps the assistive tools and curation services to understand the user needs and accordingly provide relevant pages. The tools will search the most relevant pages from the existing set of pages according to the user profile. Therefore, to retrieve more relevant information from the web, focus is now on building a detailed user profiles describing the best possible preferences of the user.

Most of the existing approaches are based on populating the user profiles automatically or semi-automatically by gathering information about users from their communication on social network websites [38, 81, 84]. For example, tagging activity on Facebook provides information about the interests of a person. However, mining social networks to build user profiles is a challenge because social website contains huge amount of irrelevant and noisy data such as ambiguous tags and words. Some other approaches consider different user behavior at different environments[39, 90]. Since the needs and preferences of a user are highly affected by her environment, they proposed a user profile with different personae. Such a profile takes care of her different preferences at different locations and provides personalized service accordingly.

One of the notable profile management schemes is proposed by Golemati et al.[40, 41]. They proposed a context-based visualization scheme to provide the best visualization properties of a web-page to a user. The scheme maintains separate ontologies for a user, her system and the document collection. The user ontology keeps track of different abilities of a user that includes visual memory, arithmetic memory, color recognition, and her educational background. The method of identifying these abilities is based on her responses to certain queries. The hardware equipment details of her system: processor, memory, graphics, and input output devices are present in the system ontology. The document ontology contains metadata: author, title, related keywords, and document category. These ontologies communicate with each other via matching rules and provide the most appropriate visualization property to the user. This visualization property is selected from several available schemes designed by considering the experience and capabilities of the user as well as his system. However, the proposed scheme is limited to provide a page according to the cognitive abilities (visual or arithmetic) of a person and neglects other previously discussed issues influencing web accessibility.

Building user profiles automatically, semi-automatically or manually has always been a challenging task. For example, taking feedback from the users to improve their profile disturb their workflow. Moreover, building profiles automatically do not consider any changes from the users. Therefore, building

user profiles automatically providing users the ability to edit their preferences seems to be an ideal solution. However, most of the existing approaches have centralized storage of the user profiles with users having little or almost no control on their profiles in terms of changing their preferences.

6.3 Our approach: Solving web inaccessibility with re-narration

The existing approaches help the curating services (such as Google) in retrieving the most relevant set of pages for the user from the existing resources. The questions still not explored completely are: How to personalize a web page to a person who is unable to access it? Can the user-profiles identify the inability of a person and provide an accessible version of the requested web-page automatically? We formulate the following research question, *"Based on the user's preferences, can we dynamically render a page by re-constructing its elements with different narratives?"*. In the thesis, we try to answer the question by proposing a re-narration approach to the web.

Re-narration is a general activity that we have been using for centuries. A teacher re-narrate a story book to her students to invoke their interests and obtain some learning. Newspapers are very good examples of re-narration where people narrate the news to others based on their interests and capability of understanding. For example, it may be hard for people to understand a government act, therefore, people try to get its narrations through some agents or learned people. Several such services are also available on the web where people, based on their context and interests, subscribe to get more suitable narrations of different topics such as news and technical issues.

We are interested to extend this re-narration approach to solve inaccessibility issues of the web. In the web context, re-narrating a web-page is to rewrite its DOM[6] elements based on the rewriting specification (Figure 6.2). Almost all the existing approaches to solve accessibility issues handled at browser level can be recognized as a syntactic restructuring of the DOM structure of the document with respect to particular needs. Examples of restructuring a page include: reading alternate text for images with the **alt** tag when a blind person visits the page; displaying text captions for a video when a deaf person visits the page; systematic replacement of colour in the document to make a document accessible to a person with colour blindness. Each such renarration, therefore, can range from a simple metadata completion as above, or can be a translation into another language, or an audio narration, or a simplification of text, or a description with other multimedia content. Therefore, for one page, multiple re-narrations can exist in any possible form of multimedia.

Currently, all the existing techniques follow a single point of architecture where the web curators are responsible for making the web accessible to everyone. The traditional approaches implement fixed rewriting strategies that will work for all the users visiting the page. Using restructuring as a model for accessibility frees the author of the page from implementing specific rules. We are looking towards diffusing the single point architecture of the web by allowing people to take responsibility of generating

[6]Document Object Model(DOM) is a language-independent representation of a document which shows interaction among its different objects.

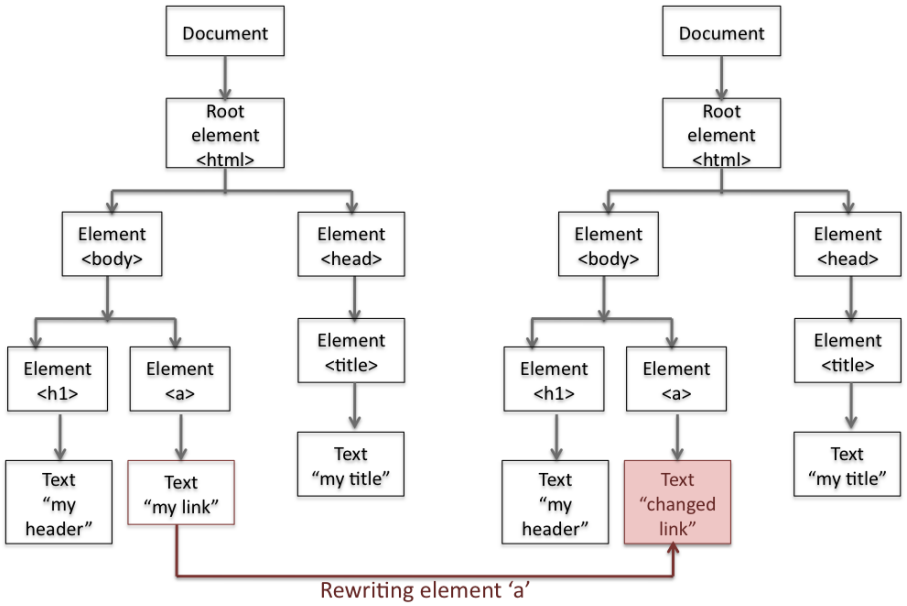

Figure 6.2 DOM of an HTML document presented as a tree structure where element 'a' is rewritten with some other text.

accessible content by re-narrating different elements of the page. Here, the re-narrators can be the page author, a user, a third party, or even a re-narration service. The motivation behind such an approach is the success of social networks where people share information with their friends, family members and community members. People are aware of the needs of their community members, friends and family members and thus may be willing to help them. Therefore, we believe that every individual has potential to make the web accessible to atleast a few small groups of acquainted people by using the re-narration approach.

6.3.1 Revisiting web accessibility issues with re-narration approach

Consider a web page of fire safety shown in Figure 6.3, authored in English, and has few images and text in it. To make this page accessible, WAI provides some guidelines as to how the page structure should be designed by the author of the page (WCAG), considering both the accessing technologies (UAAG) and rendering tools (ATAG). Let us demonstrate how to address different inaccessibility issues using this general approach of re-narrating web pages with social contributions by considering different scenarios.

Physical Disability

Let us consider a specific case of blind person, say B, visiting the page. The WAI ensures that the blind person should be able to read the page using screen-reader software and flexible navigation[24, 55, 86].

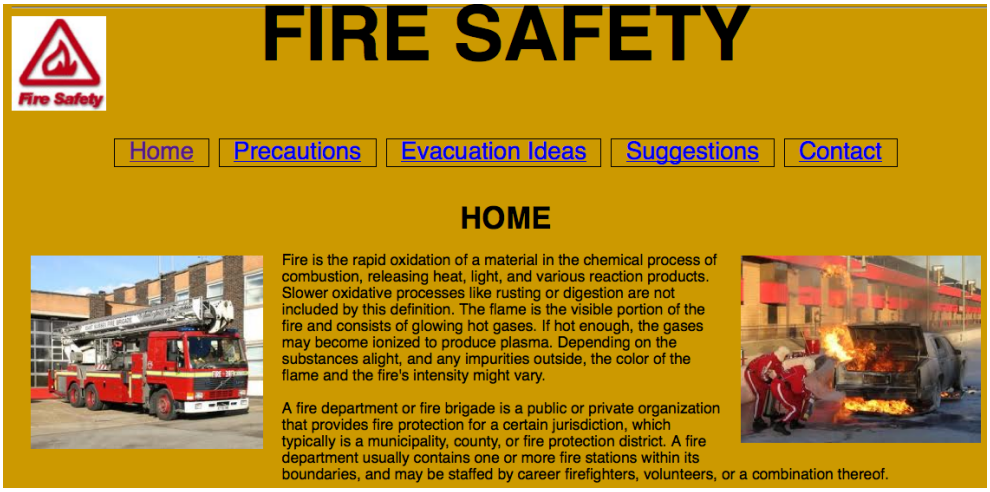

Figure 6.3 A web page on fire safety with text presented in English

The author of the fire-safety page, say P, would have to make sure that the page has clear structural information with proper headings, image captions, alternative text for image tags, and should avoid any use of elements that are difficult to be read by screen-reader (e.g., using tables for presentation is a very common "misuse").

Now, consider a case where page author violated one or more of the WAI guidelines. In this case, B might not get an appropriate page. WAI has nothing to offer to such a user. We want to explore a re-narration approach where X might be able to provide a compatible version of the page P. Instead of reading P, B can access the re-narrated page, which is a compatible version of P for B. However, the existing approaches do not explore the community based re-narration of the web completely. The approach proposed by Takagi et al.[94] to provide accessibility tags using social collaboration is a demand-based service where person B should report for inaccessibility of P to get its accessible version. Moreover, the existing profile-based approaches would help to choose one of the visualization schemes from a set of user preferred schemes, but would not create accessible content for B.

We are willing to explore something similar to the approach of Takagi making it more flexible from request-based service. There could be many X producing different compatible versions of page P by re-narrating its different elements. For example, one person can provide an audio snippet that reads out the content of page P, making it more suitable for B. Another can provide an alternative text to the image of fire-safety bus. So, B now has the choice of picking one or many from the available multiple re-narrations. This is a more general approach where everyone is contributing to solve inaccessibility problem – in a way more in the spirit of Web 3.0.

Language barrier

Suppose page P is shared with a person L who can read Hindi but not English. In this case, page P is not accessible to L because of print illiteracy. Now, when L visits P, L would understand the page if there exists a Hindi version of P. One possible way for L is to use Google Translator service and read the page. But consider the cases when L prefers to read human translations, or if Google translator does not cover some local dialects of Hindi like Braj[7], Bundeli[8]. L might prefer trusting a friend in giving her a more reliable translation of page P. Figure 6.4 shows Hindi version of the page P narrated by an acquainted person to L.

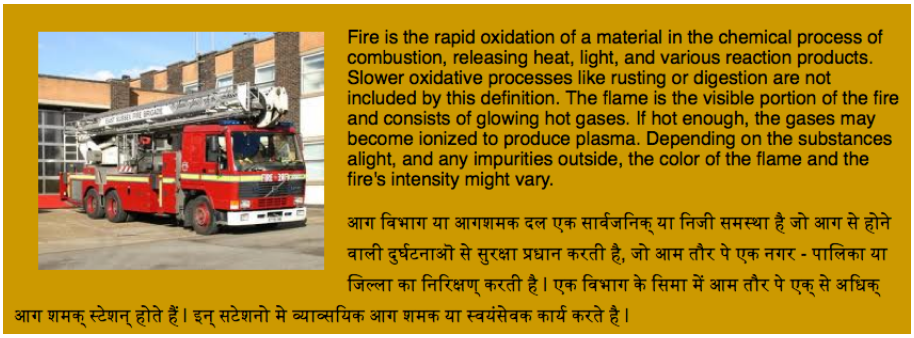

Figure 6.4 Fire safety page with one paragraph narrated in Hindi

The other ways in which the re-narration approach might address language barriers are when (a) an audio narration in Hindi is available for the English text on the page, or (b) a Hindi audio track is available for an English video, or (c) a Hindi textual commentary is available for an English video.

Social and geographical barriers

The image of fire-safety bus available on page P might not be something that person L has seen around in his town. It might be a picture of a bus service used in the US. This also contributes to print illiteracy where the context is different for a user. If the re-narrated page of P can also substitute an image of fire safety bus found in India, more specifically to the one available in hometown of L, this problem is solved (Figure 6.5). This kind of cultural re-narration is harder to achieve in language translation services and other existing approaches.

Literacy Barrier

Today, major portion of the Web is dominated by text. In countries like India, literacy rate is only about 74%[7]. Yet this does not imply that an illiterate person can not comprehend the complex issues related to government, law etc. described on the web page. In order to make the page meaningful to the illiterate person, the content can be made available in a different format that is accessible to him. For example,

[7] A Hindi dialect spoken in western Uttar Pradesh
[8] A Hindi dialect spoken in west-central Madhya Pradesh

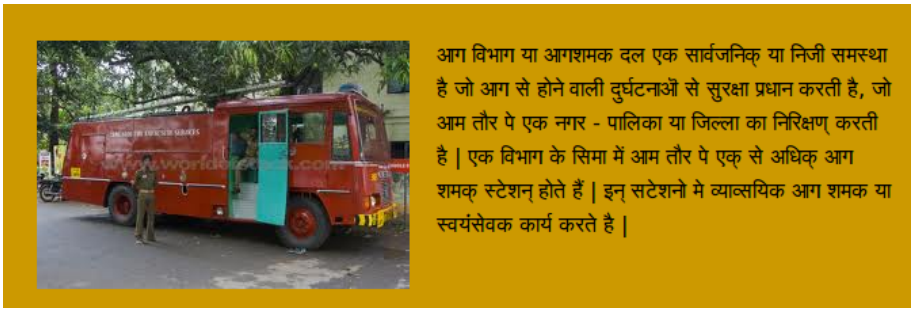

Figure 6.5 Fire safety page with image re-narrated to a local fire bus

providing oral medium in form of a video or audio might be helpful for an illiterate person. Again, considering the re-narration approach to address this issue, a Hindi-speaking but not Hindi-literate person would be able to find a suitable re-narrated version of page P.

Other accessibility issues

Apart from the language, geographical and literary barriers, there are cases when a page is not accessible because of poor writing or availability of too much technical content. The existing approaches provide dictionary meanings to difficult words of the page to help the reader. An alternative approach of providing re-narrations in form of simplifying the content might be more helpful. For example, Government documents specifying laws are not directly comprehensible to everyone visiting the page. Local versions of the page to workers, and to other readers might be more appropriate. So, a re-narration service might act as a bridge connecting technical pages with simpler narrations.

6.4 Conclusion

The goal of Designing for Social Inclusion is to render the Web that is accessible to everyone across varied abilities, age, culture and geographical locations. A web-page is accessible to a user if she is able to consume the conveyed information without any hindrance. There have been efforts to solve the web inaccessibility with different guidelines, plugins and user profiles. Furthermore, physical disabilities are traditionally given higher priority and lesser attention is given to other factors affecting web accessibility. Contextualization and localization of the web-pages according to the user preferences are still unexplored issues related to web inaccessibility.

Collaborative approach of contributing to a web page either in improving its DOM structure or its content accessibility is promising. The web becomes a much more effective medium of knowledge when users and information consumers have access to interpretations or re-narrations of content. Several rudimentary forms of re-narrations already exist today on the web as blogs, annotations on pages, bookmark recommendations, tagging, etc. Little support, however, seems to exist in the meta-data frameworks of

web pages that allows a re-narrator to target a specific group of readers, based for example, on language, location, etc. Likewise, the current architecture of the web do not explicitly support the user preferences for a particular set of re-narrations to be automatically retrieved. We are looking at a general model of the web which can solve all the inaccessibility issues uniformly. We believe re-narration approach to the web has potential to reach every problem. We, therefore, propose a re-narration based framework of the web in next chapter.

This page is intentionally left blank.

Chapter 7

Alipi: A Framework for Accessing Multi-lingual Web

7.1 Introduction

We propose Alipi, a distributed and participatory approach for re-narrating web pages for the purpose of rendering the content with better comprehensibility and accessibility. Alipi is a framework designed with the objective of enabling one set of web users, i.e., the 're-narrators' to re-narrate any web page or its element, and a second (possibly overlapping) set of users, i.e., the 'readers' who consume the web resource appropriately re-narrated to them. Alipi means print illiterate in Kannada and several other languages of the Indian subcontinent. Our framework tries to fulfil its literal meaning, and therefore, supports alternative descriptions for a webpage or parts of it via rewriting or re-narration for a given target audience by volunteers. The Alipi approach is in the spirit of Takagi et al. [94], but our approach is somewhat broader. Similar to the approach by Mirri et al. [77], we also rely on browser plugins, website-toolbars, and decentralized servers for generating renarrations of pages.

Alipi is a social collaboration framework for authoring, targeting and accessing re-narrations of web pages. The components of Alipi framework consist of a predefined set of web element attributes: a browser plugin for creating re-narrations at the re-narrator's end and for generating the re-narrated page at the reader's end. Alipi supports an architecture where semantic attributes derived from the content of page are mixed and matched with the semantic attributes of a particular reader. Additionally, re-narration is applicable to every individual element of the page. The combination of these features makes Alipi, initially designed to address print illiteracy, usable in much more general contexts. For example, using Alipi, it is possible to combine selective translation of a page with splicing of locally relevant images in order to make information accessible in a broader sense. All the issues of web inaccessibility which we discussed in Section 6.3 of Chapter 6 can be successfully solved by using our re-narration framework, Alipi.

The traditional solutions for accessibility demand that the author of webpage take responsibility for ensuring accessibility of the page. This is usually done by the author specifying a rewriting rule usually fixed as a standard across all pages, for example, providing alt tag to images. The approach used by Alipi is that these rewriting rules need not be fixed a priori. There might be multiple versions of these rewrites

for example, by a user, the page author, a third party, or even the renarration service. Fixed strategy is then a special case of the Alipi approach where only one standard re-narration is available correspoding to a page. Alipi accommodates multiple strategies for accessibility: fetching re-narrations of a page from somewhere else on the web, or restructuring a page in place based on a standard specification without fetching anything externally, or a combination of the two, where rewriting parts of a document requires fetching a re-narrated snippet from an external service.

Rest of the chapter is organized as follows: Section 7.2 defines the Alipi architecture with its three sub-systems. Alipi prototype is explained in the Section 7.3 where steps to use Alipi system are explained. In Section 7.4, we explain the procedure and the results obtained from the study conducted to test the developed prototype. Section 7.5 discusses the browser plugin developed to support Alipi approach along with an explanation of its working. Lastly, the chapter is concluded in Section 7.6.

7.2 Alipi Architecture

Alipi relies on three main subsystems: (a) a subsystem for re-narrators to create narrations, (b) a subsystem for indexing different elements of web pages to their re-narrations, and (c) a subsystem for web-page readers to display the renarrated page dynamically.

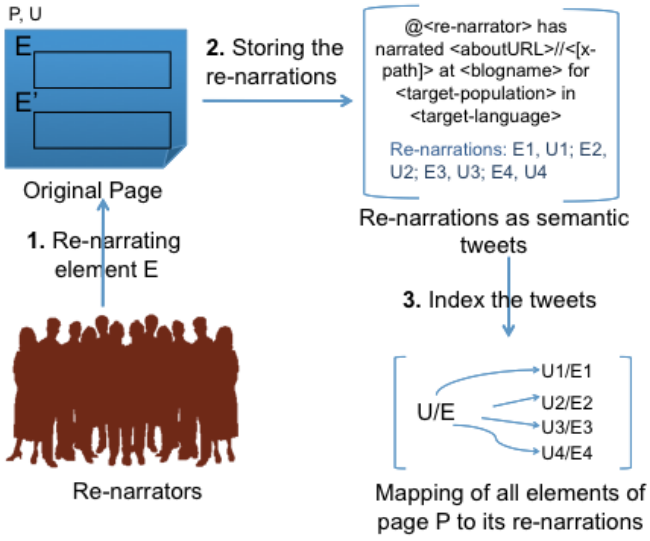

Figure 7.1 Schematic describing the approach followed by Alipi: Several re-narrators renarrate different elements of a page, P. These re-narrations are stored at different blog spaces but are maintained as an index on Alipi server.

Schematics capturing the architecture of Alipi are shown in Figures 7.1 and 7.2. In Figure 7.1, re-narrations of the web page P consisting of multiple elements (E and E') are being created and indexed.

A set of re-narrators create a set of re-narrations E1, E2, ... and so on, for the element E (arrow 1). These re-narrations exist as independent entities on the web each with its own url U1, U2, ... and so on. For example, these re-narrations can be stored at personal blog spaces of the re-narrators. Alipi requires all re-narrations to be publicly accessible pages on the Web in order to ensure a decentralised re-narration model. The decentralized re-narration model of Alipi is important for allowing users to have control of their re-narrations and to decouple documents from their re-narrations so that they are treated as regular Web pages. The information about each re-narration (such as target language and population of the re-narration) is stored as a tweet in our database (arrow 2). The tweets are indexed on the Alipi server where each element of the page P is mapped to its different re-narrations (arrow 3).

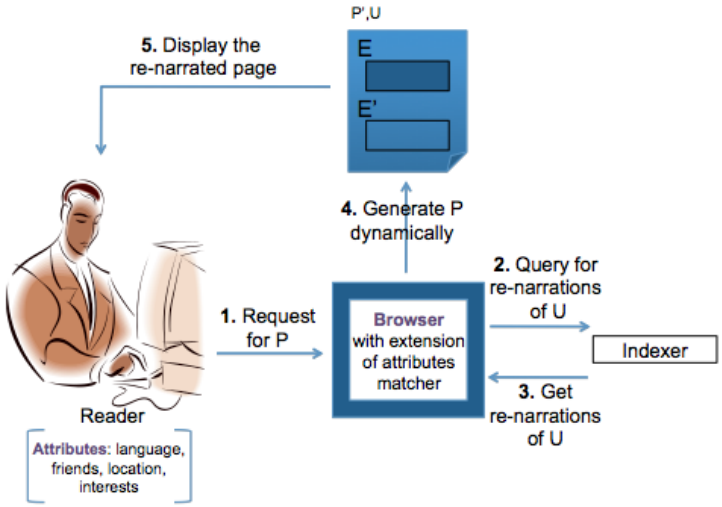

Figure 7.2 Schematic describing the generation of a renarrated web page by the browser plugin. The plugin filters some re-narrations suitable for the reader based on his preferences and then dynamically generate an accessible version of the page P.

Figure 7.2 shows how a page with possibly several re-narrations is rendered to the user consuming the page. When a user requests the page at url U (arrow 1), the user's profile containing various semantic attributes are sent to an attribute matcher. The matcher queries the indexer for the appropriate set of re-narrations of the requested page. The appropriate re-narration, chosen on the basis of the user's semantic attributes are then rendered in the user's browser as a re-narration P' of P at the same url. The architecture proposed affords flexibility in terms of implementation. The set of semantic attributes that identify a target group can belong to ontologies defined and published by the re-narrator. The indexer could leverage the semantic attributes related to the target group, e.g., language and location for efficient retrieval. The matching could be done either at the user's end, or at the index server's end, or even at a separate "matching server" depending on the application. The matching process could range from simple attribute matching to a complex set of matching between ontologies combining several

re-narrations. The generated page could be composed at the matching server and delivered to the user's browser.

7.3 Alipi Prototype

The Alipi prototype implements the core ideas of the Alipi architecture for re-narration. In the prototype implementation, the re-narration is implemented as a service. A screenshot of the server's entry page is shown in Figure 7.3a. A user visiting this service can choose a webpage for re-narration, specify the target groups and publish the re-narration at a url of her choice.

The re-narrator can either define alternative text such as translations or simplifications or provide alternative media such as audio or video according to the target audience. The re-narrator also provides meta information such as language, geographical region, nature of re-narration (translation, simplification etc.), and tags to identify the target audience. The re-narrator publishes the re-narration once it is completed. Alipi keeps track of the source, target, and language of each re-narration. Any number of re-narrations may exist for any given source page. Typically, a re-narrator will publish the re-narration at say, her publicly accessible blog. Alipi maintains a blog for those who do not have their own blog. Re-narrated posts using Alipi service are indexed on an alipi server.

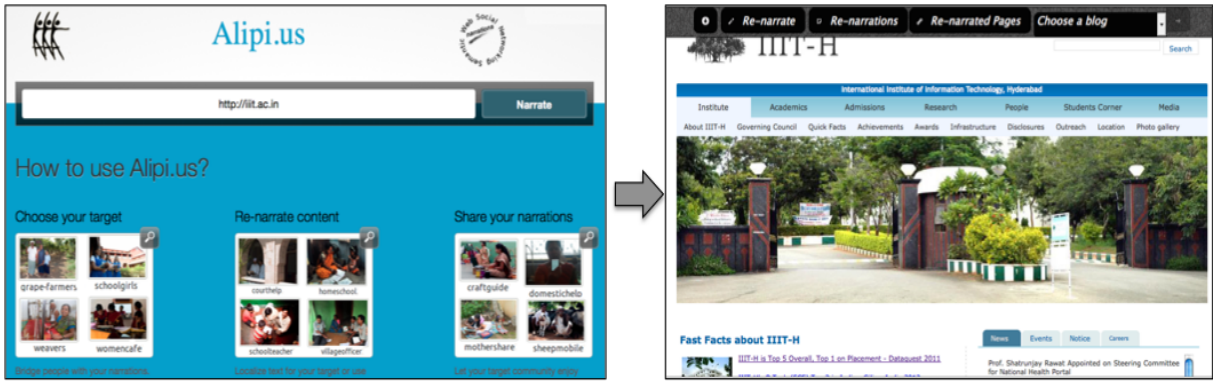

Figure 7.3 Alipi browser service: a) User types the URL in the given textbox and press 'Narrate' button. b) The typed URL will be opened in another tab with a toolbar on top of the page.

Alipi renders re-narrations by user choice. Furthermore, it can merge multiple re-narrations of a document in order to deliver the most complete re-narration. This is done by examining the xpath ids of the re-narrated elements. A user may have a locally installed browser extension for carrying out the re-narration. Prototype browser extensions for Firefox and Android has been implemented [53]. This extension also indicates the availability of re-narrations for a user requested page.

7.3.1 Steps to use Alipi prototype

Below are the steps that a user needs to follow to re-narrate a page using Alipi (refer Figure 7.3, 7.4 and 7.5).

1. Firstly, the user needs to visit the URL: http://alipi.us. She then types the URL of the web-page she is interested to re-narrate such as http://iiit.ac.in, along with http, as shown in Figure 7.3a. After pressing the button 'Narrate', the requested page will be opened in next tab of the browser with a toolbar on top of the page (Figure 7.3b). This toolbar is provided by Alipi service which will enable the user to rewrite the page.

2. User needs to click on the 'Re-narrate' button to write her narration of the page. She can then select any element of the page such as an image, paragraph, hyperlinks or any other DOM element of the page. On clicking the individual element of the DOM, a window will pop up where the user can write her narration corresponding to the element (Figure 7.4a). For providing a video or image or audio re-narration to the page, the pop-up window asks for the source URL of the re-narration.

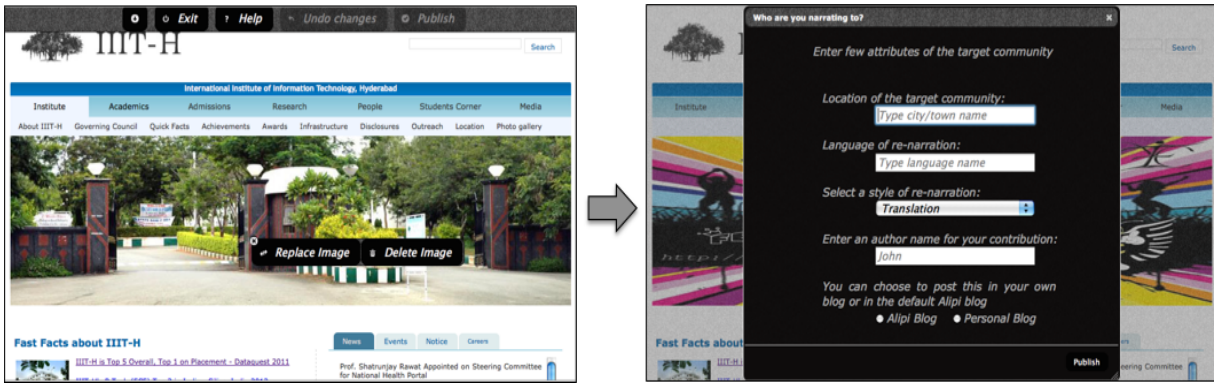

Figure 7.4 Steps to re-narrate a page. a) After clicking on the re-narrate button, user can click on any element of the page to re-narrate. If the user clicks on the image, a window to re-narrate the image pops up. b) To publish her re-narration, the user is required to fill the necessary details about it.

3. After re-narrating, the user needs to click on the 'Save changes' button. The re-narration is then saved and the user is re-directed to the original page to write more re-narrations for other elements of the page.

4. Finally, to publish the narration, the user needs to press the 'Publish' button of the Alipi toolbar. While publishing, the user is required to mention certain details about the re-narration such as the target population and location of the re-narration, re-narration language, blog-space where the re-narration should be published (Figure 7.4b).

5. Once the content is published, the user can see it along with other existing re-narrations (if available for the page) on clicking the 'Re-narrations' button.

To check the available re-narrations for a web-page, firstly, the reader needs to follow the above mentioned first step. Then, she needs to click on the 'Re-narrations' button and select one suitable language from the available list of the re-narrations (Figure 7.5a). For the chosen language, all those elements of the page will be rewritten to the re-narrated content for which the re-narrations are available. In Figure 7.5b, the image is changed to the re-narrated image.

Figure 7.5 Steps followed by a user to see a re-narrated page: a) User needs to choose language of the re-narration from the list after clicking on 'Re-narrations' button. b) On choosing the option 'others', image of the page is replaced with the re-narrated image.

7.4 User study of the Alipi prototype

Study Objective

Since this was our first laboratory study of the project, we were interested to know the experience of people with our developed prototype based on the Alipi architecture. We were also interested to study different attributes related to both sets of people defined by Alipi, namely, readers and re-narrators of the web-pages.

Participants

We conducted a small scale study of Alipi in the R&D showcase of our university with 70 participants (male=45, female=25). These participants were Engineering students of different colleges, faculty members, and professionals from different companies. The age of the participants ranged from 18 to 45 years with average age 22 years. Thirty percent of the participants were active on blogs, forums or YouTube for technical discussions and recreation purposes (such as sharing poetries, music) while the rest were not active but were casual users of these sources. However, all of them considered such applications as

ideal platforms for open discussions allowing them to share their knowledge and thoughts.

Procedure
The study was conducted in English. We were available to help the participants throughout the study. The study was conducted in three phases: In the first phase, demographic information of the participants (name, age and profession) was taken. We asked a few questions regarding their experience with other available services like: YouTube, forums and blog (Appendix A.4). We explained our system to everyone and gave them an insight on how to use it as a re-narrator and reader. In the second phase, we asked each participant to re-narrate a page of their interest or to see a re-narrated page. To demonstrate the working of our tool, we took two web-pages: IIIT-Hyderabad homepage with URL as http://www.iiit.ac.in/ and a page on Indian culture with URL as http://www.culturalindia.net. However, participants were free to re-narrate any page of their interest. Participants were encouraged to think aloud while performing the task. The last phase was dedicated to one-on-one interview with the participants where we informally discussed with them about their experience of our system. In the interviews, we asked the participants some questions covering different aspects of our system (Appendix A.4).

7.4.1 Study Findings

Most of the participants tried the tool with their college and company websites. Our selected two websites (IIIT-Hyderabad homepage and Indian culture page) were used for the cases where college websites were not working or where the user was open to any web-page. The re-narrations were generated in text, video and image. No participant chose audio as a medium of re-narration. Participants appreciated the idea of re-narrating pages to enhance the accessibility of the web pages and were very much interested in using our system frequently. We present our findings of the study as follows:

1. *Participants were interested in playing both the roles of a re-narrator as well as a reader.*
   We obtained mixed views for the role people would like to play: the role of a re-narrator or a reader. Participants were excited to re-narrate a page for the reasons of sharing their thoughts and feeling of control over the web-page content. However, the choice of role for them was dependent upon the information present on the web-page. According to the participants, if the domain of the web-page was known to them, they would re-narrate the page and share their knowledge with others and if it was unknown to them, they would prefer to see the re-narrations of others. However, for topics like politics, cricket match updates and news headlines, every participant wanted to re-narrate as well as read the available re-narrations: “I surely would want to see the comments of my friends about Dhoni’s performance in last match.”

2. *As a re-narrator, people preferred generating text based re-narrations to video and audio re-narrations.*
   Sixty percent of the participants preferred to generate text based re-narrations because they found writing text is easy and less time taking. Twenty percent of them were willing to generate video

narrations also, if provided with the required set-up in a user friendly way. To describe this, one participant said, "I would love to post a video explaining a topic, but I don't want me to be involved in getting a camera, setting it up and talking to myself about the stuff. Maybe I will appreciate an online canvas or a realtime board for the same!" For forty percent of the participants, the choice of medium was dependent upon the web-page topic and they were willing to use any medium which could convey their intention in a most easy to understand and concise manner.

3. *As a reader, people preferred re-narrations in mixed media.*
   Eighty percent of the participants preferred to check a combination of multiple forms (text/ video/ audio/ image) of re-narrations for a page. The reason is that a mixture of different media would express the perceptions of people in the best possible way. Sixty percent of the participants were interested in checking the video re-narrations first, followed by text and audio because participants considered video as more descriptive and interesting medium of communication: "Whenever I want to learn a new tool, I watch a YouTube video to learn its basics." They mentioned that for situations like low bandwidth and other degrading factors, they will prefer text based re-narrations. Only two participants mentioned to choose audio re-narrations for the page while others found audio re-narrations inappropriate: "It is very irritating to listen to someone about a topic without watching him." Image based re-narratives were welcomed by everyone but with a combination of some other re-narrations available in text or video.

4. *Majority of the participants were willing to generate re-narrations for their friends.*
   Ninety percent of the participants showed their interest in re-narrating pages for their friends such as their college and school friends group. Sixty percent of them found our system useful in helping their family members by increasing their browsing experience and making the page more comprehensible: "My mother is very much interested in cooking recipes but she is not much proficient in English. If I translate my favorite dishes to her, she can cook more for me." Ten percent of the participants mentioned farmers and illiterate people as their target groups for their re-narrations and were interested to narrate pages specifically in videos for them.

5. *Most of the participants wanted to see re-narrations from known people.*
   Eighty four percent of the participants were interested in seeing re-narrations from their friends and relatives and were less inclined towards considering re-narrations of unknown people: "My friends know me better. Therefore, they will re-narrate the pages in the best possible way for me." Only 22% of the participants mentioned their interest in seeing re-narrations from unknown re-narrators and wanted to search the best re-narrators themselves depending upon the topic: "I usually follow several experts on technical discussion forums for coding in different languages. Similarly, I would like to identify experts for the page of my interest by reading their available narrations."

6. *Participants were interested in generating as well as checking the re-narrations in their local languages.*

   While re-narrating pages, majority of the participants were inclined towards developing webpages in their local languages because they found language as the biggest barrier to the web accessibility. Seventy eight percent of the participants showed their interest in generating and referring re-narrations in their local language whereas the rest twenty two percent were neutral to the language preferences but were more focussed at the content of re-narrations.

7. *Participants found our system very useful to share information.*

   Participants found our system useful to discuss and share information about various topics ranging from sharing class notes with colleagues, to sharing comments on discussion forums and news updates, to helping specific crowd like parents and farmers, to helping physically disabled people. One college student stated, "I guess it will be really useful when notes made by our class topper can be used by us!" and another participant said, "I can mark stuffs as cool while surfing online hacking stuff helping new hacker to filter out things faster!"

   On asking whether our system would lure them to create some noisy data for fun and enjoyment, majority of the participants disagreed to it. They appreciated the system for learning and sharing purposes. One participant said "If I were to write garbage then I already have Facebook and Orkut, I will not use this."

8. *Participants found the interface design of our system non-intuitive and uneasy to follow.*

   The design of our system demotivated some users to re-narrate pages. Since our system requires a sequence of steps to re-narrate a page, most of the people found it very complex and prefered being a reader rather than a re-narrator: "The tool should be as simple as liking a facebook page, writing recommendations on LinkedIn, pressing thumps up on youtube videos." Another participant said, "Role of the reader is more lucrative at this point, first because I am lazy and second because I don't know an easy way to do re-narration." Participants found the re-narration steps non-intuitive to follow. For example, the 'Save' button to save a re-narration gave participants an impression that their narrations were published and therefore, the next intuitive step for them was to click 'Exit' button. However, to publish a re-narration, our system requires clicking 'Publish' button as the next step, which confused them tremendously.

   Every participant suggested to develop a better GUI where the steps would be more intuitive and easy to accomplish. Some of the other suggestions in the direction of improving our tool were as follows: to merge 'Save' and 'Publish' button together for reducing confusion of the users; and to give proper feedbacks to the users in order to guide them the usage of system. One participant mentioned that our system should also allow people to write text on images where the re-narrations would serve as an alternate text or caption, which is not present in the current prototype. Five participants suggested to add features such as providing 'Like' button on each re-narrations to increase their credibility.

9. *Majority of the participants were interested in seeing an automatically transformed page with the best set of re-narrations suitable for them.*
   Sixty percent of the participants mentioned their interest of seeing a re-narrated page converted automatically by considering their profiles and interests; twenty two percent mentioned to manually select the best suited re-narrations for them. The rest eighteen percent were neutral about any such automatic transformation of the page as they wanted to experience the system for a few days to decide their preference for manual or automatic selection of the re-narrations. Thirty percent of the participants wanted an editable profile to regularly update their interests and selection priorities to improve ranking of the re-narrations for their queried pages.

7.4.2 Discussions

While conducting the study, we faced certain issues with our system. Our prototype did not work well with Wikipedia pages, news web-pages and other dynamic pages like http://cricinfo.com. Additionally, the alipi toolbar was not appearing properly for certain universities' web-pages, e.g., webapge of IIIT Bhubaneswar, http://www.iiit-bh.ac.in. However, we did not allow these issues to affect our study. Whenever any such issues were encountered with the requested URL, we asked the participants to experience the tool with our selected web-pages.

To make the task of re-narration easier for the users and to meet the bigger goal of Alipi, i.e., to dynamically render web-pages according to the user requirements, we have developed a browser plugin. The plugin stores the user preferences as user profiles and allows users to re-narrate any web-page. In the next section, we will discuss the plugin in detail.

7.5 Alipi prototype as a browser plugin

We are currently working on a Firefox plugin to support our Alipi framework. The plugin by passes the URL http://alipi.us and enables every page for re-narration. The purpose of plugin development is to provide an easy solution for the readers to access the re-narrations. Figure 7.6 shows a snapshot of the Alipi plugin toolbar which appears in the browser after its installation. In the Figure, menu 'Renaration' shows the number of suitable renarrations available for the user; Menu 'Author' shows details of the current re-narration (author and re-narrated language) recommended by the plugin; button 'Renarate' allows the user to renarrate the current web page and the button 'Original' allows the user to see the original web-page; menu 'Settings' contains different options to maintain the user profile. Further details about the plugin development can be found in the technical report by F. Boudinet et al. [36].

The plugin works in the following way for the three subsystems of Alipi:

1. At the reader end, plugin establishes a connection with the Alipi indexer server and retrieves all the available re-narrations for the queried page. The plugin filters some of the most suitable re-

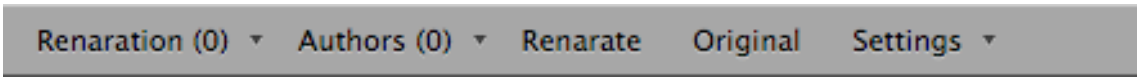

Figure 7.6 A snapshot of the Alipi plugin toolbar showing different options of menus and buttons for the users.

narrations for the reader. The selection is based on her profile along with some rules defining how her specified preferences should be considered. We will explain the user profile and filtration algorithm of the plugin in the next section. In the current version of the plugin, all the filtered re-narrations are listed under the 'Renaration' menu of the toolbar and the reader needs to choose one from the list (Figure 7.6). But we are working to modify the plugin such that the original page gets re-narrated on-the-fly for the highly matched set of re-narrations available for unique sections of the page, while the remaining filtered re-narrations will be listed under the 'Renaration' menu of the plugin toolbar. The plugin provides necessary details of all the re-narrations available for the page such as the re-narration author, language of the re-narration, and the list of re-narrations.

2. For a re-narrator, the plugin provides a 'Renarate' button to re-narrate any page (Figure 7.6). The user needs to go through the same steps from step 2 onwards as explained in Section 7.3.1. Corresponding to each re-narration, following information will be stored as a semantic tweet on the Alipi server: re-narration target information (location, language, community), blog URL where the re-narration is available and re-narrator details (name, specialization).

3. Indexing of the re-narrations available for the web pages remains same as explained earlier. The available tweets are mapped for each section of the web-page, which are then fetched for every request to the page.

7.5.1 Maintaining user profile with Alipi plugin

The plugin maintains a user profile under 'Settings' menu of the toolbar (Figure 7.7a) and the profiles are stored as a JSON[1] file on Alipi server. The profile is editable and contains the following attributes: Language as read-and-write and listen-only; Disability as visual, hearing, none; Friends, relatives and Interests (Figure 7.7b). For example, if a person X knows three languages: English, Hindi and Telugu and he is proficient in both English and Hindi (read and write ability) but partially understands Telugu (Listen-only). Such information about language proficiency is useful to filter out the suitable re-narrations for the user. For example, a text based re-narraton available in Telugu language is not useful for X. However, an audio re-narration in Telugu may be suitable for him. In addition to this, the language distinction makes a lot of sense with disabilities. For example, if X is blind, he may still

[1]JSON is Javascript Object Notation. It is a text based simple notation of data structures to enhance human-readability of the data.

be able to listen re-narrations but can not read the web-page (considering any other accessibility tools are not installed in X's computer). Here, language type 'listen' should be considered by the plugin to suggest best re-narration to X.

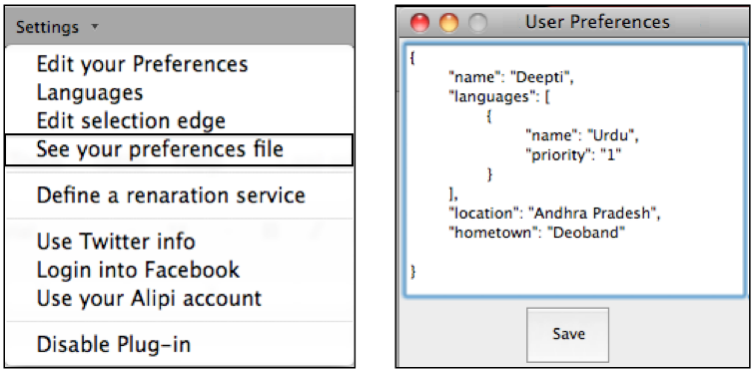

Figure 7.7 Maintaining user profile with Alipi plugin a) 'Settings' menu overview: The user can edit her information, disable/enable the plug-in and import her profile from Twitter or Facebook. b) User can edit her preference file by writing her interests in JSON format.

The user needs to fill his profile in the 'See your preference file' manually. The plugin also allows the users to export their profile from Facebook or Twitter. If a user edits her preference file, the changes are reflected back to her JSON file maintained at the Alipi server. We are currently working on providing an easy user interface to edit profiles and on extending the attributes of user profile to include more details of the user.

7.5.2 Algorithm to recommend suitable re-narrations to the reader

The plugin runs a selection algorithm to filter out the most suitable re-narrations for the reader by checking the available re-narrations against her profile. Below are the rules of recommending a re-narration to a reader A.

1. Target language of the re-narration should match one of the specified language in the Language list of A's profile. Languages mentioned as 'read-and-write' are given higher priority than 'listen-only' languages. For example, if language type of the available re-narration is read-and-write, the re-narration in any medium (text, audio, video) can be recommended to A. However, if any re-narration is not available for her language preferences under read-and-write category, then plugin will recommend the available video or audio re-narrations for her prefered languages under listen-only category.

2. If a person has some disability then medium of the re-narration is chosen accordingly. For example, plugin will recommend text based re-narrations to a deaf person for the languages men-

tioned under read-and-write category and will not consider recommending audio or video based re-narrations.

3. Re-narrations are filtered based on the given list of friends and relatives of A. Those re-narrations are selected for which the re-narrator is either her friend or relative.

4. Those re-narrations are selected for which the target location matches with any of the locations mentioned in her profile.

5. Most recent re-narrations are then suggested to the reader. Freshness of the re-narrations is maintained by storing a timestamp corresponding to each re-narration.

In the current version of the plugin, interests of a person and disabilities (like learning & cognitive and motor related) are not considered to rank the re-narrations. Moreover, no priority is given to the location list of the person, which may contain attributes like hometown, visited places, places of education, current location and further more. In future, we will extend our algorithm to include all these factors as part of the selection algorithm. In addition to this, we will also incorporate trust level of friends to rank the re-narrations. The trust level may be helpful to decide filtering of re-narrations based on the expertise of re-narrators corresponding to her different interests. For example, if a person is interested in scientific articles then he may not consider his Facebook friends, rather he may prefer his LinkedIn connections. For recreation activities, he may prefer his Facebook friends.

7.6 Conclusion and Future Work

In this chapter, we have presented Alipi, a framework that supports defining accessibility in a larger context. The Alipi framework emphasises re-narration as a general approach to address accessibility over the Web. Furthermore, the decentralisation and multiplicity of the re-narrations eliminate the top-down, normative approach of WCAG guidelines. Alipi also enables re-narration communities to grow around specific needs as experienced and articulated by the communities and its accessibility enablers without global norms of what accessible content ought to look like.

In the chapter, we have discussed how three subsystems of Alipi work to achieve the goal of re-narration. We conducted a study to test the feasibility of Alipi approach and its acceptance by people. We received encouraging response from the participants as they were thrilled by the concept of re-narrating web in their own content. We also encountered some issues while conducting the experiment, on which we are working. Dynamically generated web-pages is difficult to re-narrate because of their varying DOM structure. Alipi needs to store DOM path of each re-narrated element of a page on Alipi server and refer it to fetch the re-narrations available for the page at reader's end. We have also explained the working of a plugin developed to make the task of re-narration easier. The plugin is still in the stage of development. We are interested to incorporate different attributes in the user profile of the plugin such as dynamic re-narration of the page with most suitable set of re-narrations, considering user interests and

other disability factors to recommend suitable re-narrations to the user. The current version of the plugin could retrieve user profiles only from Facebook and Twitter. In future, we look forward to connecting Alipi plugin with other existing profiels, to ease the work of users in populating their Alipi profile.

Several interesting technical questions have emerged as we embark on developing Alipi from a prototype to a more robust implementation and testing it with sizeable communities over the web. To cite just a few examples, what could be a metric for matching or comparing the relatedness of two re-narrations? What optimizations are possible in the indexing and delivery of the matching pages? What are the security implications of the architecture? Finally, in the proposed Alipi architecture, we plan to build on rich ontological structures shared across social networks created in a distributed, de-centralised manner, used with browsers extensions and web services. Thus, we foresee Alipi leveraging the Semantic Web in a comprehensive way. From a social perspective, it would also be interesting to study formally how communities share and evolve around re-narrations and what issues could emerge within these re-narrations.

Acknowledgement: Alipi is a collaborative work done with the team of Dr. T.B. Dinesh, founder of Janastu, Bangalore. This work is accomplished with the regular discussions and help of several people. We thank Arvind, Ajay and Shalini of Servelots for providing technical help in developing the system; S. Uskudarli, Subramanya Sastry for their research support and initiatives; Florian Boudinet, Kartikey Vyas and Deep Dixit for their help in developing browser plugin.

Chapter 8

Conclusion

8.1 Research Contributions

This research contributes to an understanding of developing web as a personalized assistant for the users by utilizing collective human intelligence. We created and evaluated new approaches, *Power of Friends*, and *uPick* that offer engaging ways of collecting knowledge from social networks and the online documents respectively. We also present an effective and uniform way of creating a multi-lingual web by providing a framework, *Alipi*, which enables the web users to rewrite web-pages. The main contributions of the thesis are enumerated below:

1. We reviewed the existing techniques to generate community knowledge from social networks and found certain weaknesses in the existing approaches such as lack of engagement in the given task. Since the community knowledge is scattered throughout a network of people, automated techniques fail to collect them. Moreover, the attempts made in the direction of developing human computational games to perform this task suffer in maintaining the trade off between the accuracy of the data generated and the fun element of the game. As a solution to which, we presented a new approach to the Friendsourcing games called as *Crowd Consensus*.

2. We evaluated our proposed Crowd Consensus approach by designing a human computational game called *Power of Friends*. The empirical studies of the game showed that our proposed approach has the potential to elicit truthful responses and to entice users for frequent playing.

3. We reviewed the existing automated systems to generate objective facts about named entities from the documents and found certain issues with them. Most prominent among them are dependency of the system on external resources like Wikipedia and the poor scalability of the system on different domains. As a solution to generate named entity relations, we explored the feasibility of combining automated techniques with human intelligence and proposed a new approach called as *uPick*. The approach invites human participation to filter out the system generated objective facts from a given document. We illustrated how our approach could provide an engaging environment for humans to perform this task of knowledge acquisition by conducting a user study.

4. We explored the existing systems and guidelines to make the web documents available in multiple languages accessible and inclusive for everyone. We found that inaccessibility due to physical disabilites has been given highest priority over other factors influencing the web accessibility. We worked on a framework called *Alipi*, which provides a uniform solution of re-narrating the web content to all the factors affecting accessibility. We conducted a user study to test the experience of people with a prototype developed to realize Alipi. People found the system very useful for several applications and were interested to re-write various pages for their different friend groups.

5. We developed a browser plugin to provide re-narration service to the Alipi framework. The plugin has a user profile and a re-narration selection algorithm based on which the best suitable re-narrations from the available re-narrations will be filtered and recommended to the user.

8.2 Future Directions

The research has contributed to knowledge acquisition and web accessibility literature along with certain future directions. Here, we describe the possiblities to extend our work to other interesting research problems. Below we mention a few research questions raised from our attempts while building systems for knowledge acquisition and accessible web.

1. Can the proposed Crowd Consensus framework be used to reduce the number of iterations for crowdsourcing tasks? To answer this question, feasibility of the proposed framework should be tested with crowdsourcing tasks on platforms like Amazon Mechanical Turk (AMT). If the framework produces good results, a mathematical model can be developed to illustrate reduction in number of iterations in collecting responses from the contributors.

2. Our proposed approach of Crowd Consensus does not pose direct questions to every community member and consider their belief in calculating final truth value related to the given question. Our approach finds some similarity with belief modalities as we are interested in collecting a third level information from each community member. Using the belief modality, can we develop a mathematical model to prove that the answer generated by using the proposed approach will be accurate and to determine different conditions where the accuracy may deviate?

3. Can the proposed uPick approach be useful in enhancing the experience of students while reading textbooks? This can be achieved by developing human computational games based on uPick approach, which provide interesting ways of learning as well as are fun to play.

4. How to check the relatedness of a re-narration (generated with Alipi tool) with the original document as well as with other available re-narrations for the same web-page? This may be achieved by using summarization techniques where the keywords of each re-narration are matched with the original document and with other available re-narrations.

5. How to check the credibility of a re-narration to filter the noisy re-narrations and to rank the useful ones higher than the others? To perform this, public voting for each re-narration may be considered by providing a simple 'Like' button as available on YouTube and other discussion forums. Moreover, to filter out the noisy re-narrations, techniques of content filtering by using external vocabularies and emotion/sarcasm detection may be useful.

6. How can we improve the re-narration selection algorithm used in Alipi plugin to make it more effective and robust? How the algorithm should be developed so that it also considers the rapidly growing online communities, every local dialect spoken in different geographical locations and the nearby areas of user prefered regions? The vicinity of user mentioned regions may be defined by considering parameters like language such as if the same language is spoken in the nearby areas of mentioned place, then re-narrations targeted for those areas can be recommended to the user. Different such factors need to be incorporated in the algorithm to generate more suitable re-narrations to the users.

8.3 The last words...

In the past few years, we have seen the enormous use of the web by people of different demographics with different requirements and abilities. Along with this, there arises the need of greatly broadening reach of the web so as to fulfil the overall aim of web to share the knowledge with everyone. Most of the difficulties with the huge usage of the web is due to the bulk of available information resulting in degrading the browsing and searching experience of the web users. Access of information poses challenge to locate the needs of people and present them the information in a more compatible and readable way.

This thesis is an effort to make the web useful and comprehensible for everyone. We have explored the possibilties of utilizing the human intelligence towards solving the issues of the web. In the same direction, we have made our efforts in generating the hidden knowledge from the web and to make the web accessible to everyone. We presented human computational games to generate objective facts and community beliefs from the web documents. In addition to this, we also presented a framework to make the web accessible by re-narrating the web content. Results of the conducted studies showed that the proposed schemes offer an engaging way of accomplishing the task while maintaining the quality of the data generated. We hope that our work will provoke thinking of the researchers to advance the discourse of the web.

This page is intentionally left blank.

Appendix A

Feedback Forms of the conducted User Studies

A.1 Paper-based study of Crowd Consensus framework

Figure A.1 is a snapshot of the paper-based study performed to test our Crowd Consensus Framework. We formulated different communities and separate forms were created for each community. Each community member was asked to fill one such form related to her community.

1. From how long do you know each other?
2. What are the types of activities in which your friend circle is involved?
3. How will you rank the bonding of people in your community? (Low/moderate/strong/very strong)
4. How do you each other preferences: by personal interaction/social networks or anything else?
5. What do you like about the game?
6. Did you find the game challenging? Why?
7. Would you like to play such games for friends frequently?
8. How was your overall experience with the game?
9. Any other comments or suggestions.

Power Of Friends

Hello all,
Welcome to the guessing game about your friends.
Following is your friend circles:

			
Kalika	Zarreen	Sunny	Sourabh
			
Sonali	Deepti	Sonal	Bhupendra

Following are few of the interesting facts about them. Let us see how many of us still remember each other habits and likings. ☺ YOU CANNOT FORGET ANY ONE OF "US". ☺

Note: Make your judgment only for the friends that are mentioned above. Do not include others.

Do you know?	Do you think that this is true? (Yes/No)	Do your friends also think the same? (Yes/No) Say yes if more than 6 friends also think the same.	How many of your friends think that this is true? (Excluding YOU) Tell me in number. 
Sonali loves watching romantic movies.			
Zarreen likes collecting images of couples.			
Sunny loves flirting.			
Sonal likes cooking.			
Kalika loves eating sweets.			
Sourabh likes reading novels.			

Figure A.1 Paper-based study of Crowd Consensus approach

A.2 Questionnaire of study conducted for Power of Friends game

Two sets of questions were asked from the participants: one before and another after the task. Below are the questions which were asked from the participants before playing the Power of Friends game.

Name:
Gender:
Age:
Profession:
Computer experience:

1. Do you play games? indoor/outdoor/online?
2. Which games do you play?
3. What type of game do you prefer: Challenging/scoring?
4. What do you like about the games? why do you play them?
5. How much time do you spent in playing? daily/weekly?
6. Are you on some social networking website? (Facebook, Twitter)
7. What do you like about it?
8. Do you play games available of such websites? Examples?

Below are the questions which were asked from the participants once they played the game (post-study questions).

1. Do you find the game boring/frustrating/interesting/neutral?
2. Would do you like about the game?
3. How do you find this game? Too simple/simple/neutral/challenging?
4. How did you proceed while answering questions: Is it your opinion or your friends?
5. Is there any difference in both the opinions? Where? What can be the reason?
6. Would you like to play the game again? How frequently?
7. Is this game helping you to know your friends?
8. Would you like to suggest it to your friends?
9. How do you think the game can be made more interesting?

A.3 Questionnaire of study conducted for Testing uPick

Questionnaire was consisted of two sets of questions: one set was given before giving task to the user and another set was given once the user had completed the task. Below are the questions which were asked from the participants before giving them the task of filtering named entity relations.

Name:
Gender:
Age:
Profession:
Computer experience:

1. Do you read online documents? For what purpose?

2. Do you find the way of presentation as useful medium of learning? Why/why not?

3. If not, how do you think it can be improved?

4. Do you play online games? What sort of?

Below are the post-study questions.

1. Do you find the application boring/frustrating/interesting/neutral?

2. Would you like to read documents when presented in this manner? Why/why not?

3. Is this application helping you to remember facts related to documents?

4. What if such a game is provided after each chapter of your book?

5. What about presenting the game after each paragraph, if the document is lengthy?

6. How do you think this task can be made more interesting? Can you think of any game related to this activity?

7. What do you like about the application?

A.4 Questionnaire of study conducted for Alipi system

Our questionnaire was in two sets: one set of questions was asked before giving the task to the user and another set was given once the user had completed the task. Following are the questions that were asked before participants experienced Alipi:

Name:
Gender:
Age:
Profession:
Computer experience:

1. Do you read online documents? For what purpose? What sort of?
2. Do you face any issue(s) while accessing the web? Mention some examples.
3. If answer to question 2 is yes, then how do you think that this problem can be solved?
4. What do you think about sharing thoughts on YouTube, blogs, discussion forums or any other such medium? Have you ever participated in any such discussion? What motivates you there to write?

Below are the post-study questions.

1. Do you find the system boring/ frustrating/interesting/neutral? What do you like about this system?
2. What is more exciting: re-narrating a page (Re-narrator), or seeing a re-narrated page (consumer)? Why?
3. Which medium would you prefer to re-narrate a page: text/audio/video/image/mixed? Why?
4. Which medium would you prefer to read on a re-narrated page: text/audio/video/image/mixed? Why?
5. Which re-narrations do you find useful: one from your friend, or from an unknown?
6. Would you like to have an algorithm of selecting a re-narration automatically?
7. What would be your target group of users for re-narration? Why?
8. How do you think this task can be made more interesting?

Appendix B

Project Resources

B.1 Project Resources

Slideshow of the talks delivered at different conferences and other venues are available at: http://www.slideshare.net/deepticomputer/.

Posters and publications can be accessed from my homepage: http://pascal.iiit.ac.in/ deepti.aggarwal/.

All the prototypes developed are available at the following URLs:

System	Webpage links
Alipi	http://alipi.us/
Alipi plugin (download)	http://pascal.iiit.ac.in/ deepti.aggarwal/alipi.xpi
Power of Friends	http://pascal.iiit.ac.in/ rohit/pof/
Deomgraphy based ATMs	http://pascal.iiit.ac.in/ deepti.aggarwal/atm/index.html

Table B.1 Online Availability of the Developed Prototypes

Bibliography

[1] Amazon mechanical turk. https://www.mturk.com/mturk/welcome. Last visited September 2012.

[2] Dbpedia datasets. http://wiki.dbpedia.org/Datasets. Last visited September 2012.

[3] Explicit and tacit knowledge. http://www.cognitivedesignsolutions.com/KM/ExplicitTacit.htm. Last visited September 2012.

[4] Factors influencing web accessibility. http://www.w3.org/WAI/bcase/. Last visited Jan. 2013.

[5] Games with a purpose. http://gwap.com. Last visited November 2011.

[6] Global internet usage (wikipedia). http://en.wikipedia.org/wiki/Global_Internet_usage. Last visited March 2012.

[7] Government of India, Census 2011: Literacy in India. http://www.census2011.co.in/literacy.php. Last visited March 2012.

[8] Internet use across age groups. http://visual.ly/internet-use-across-age-groups. Last visited Feb. 2013.

[9] Internet world stats. www.internetworldstats.com/. last accessed Mar 2012.

[10] Internet world stats: Usage and population statistics. http://www.internetworldstats.com/emarketing.htm. Last visited Feb. 2013.

[11] List of languages by number of native speakers (wikipedia). http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers. Last visited March 2012.

[12] Looking glass self(wikipedia). Last visited Dec. 2012.

[13] Mindnet. http://research.microsoft.com/apps/pubs/default.aspx?id=69647. Last visited March 2012.

[14] Power of 10. http://en.wikipedia.org/wiki/Power_of_10. Last visited September 2012.

[15] Social inclusion. http://www.w3.org/WAI/users/Overview.html. Last visited October 2012.

[16] Stanford pos tagger. http://nlp.stanford.edu/software/tagger.shtml. Last visited March 2012.

[17] User generated content. http://en.wikipedia.org/wiki/Usergenerated_content. Last visited November 2011.

[18] J. S. Aitken. Learning information extraction rules: An inductive logic programming approach. In *ECAI*, pages 355–359, 2002.

[19] M. S. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz. Personalization via friendsourcing. *ACM Trans. Comput.-Hum. Interact.*, 17(2):6:1–6:28, May 2008.

[20] V. Bertacco. Humans for eda and eda for humans. In *Proceedings of the 49th Annual Design Automation Conference*, DAC '12, pages 729–733, New York, NY, USA, 2012. ACM.

[21] Y. Borodin, J. P. Bigham, A. Stent, and I. V. Ramakrishnan. Towards one world web with hearsay3. In Y. Yesilada and D. Sloan, editors, *W4A*, ACM International Conference Proceeding Series, pages 130–131. ACM, 2008.

[22] S. Brin. Extracting patterns and relations from the world wide web. In *Selected papers from the International Workshop on The World Wide Web and Databases*, WebDB '98, pages 172–183, London, UK, UK, 1998. Springer-Verlag.

[23] J. Cameron and W. D. Pierce. The debate about rewards and intrinsic motivation: Protests and accusations do not alter the results. In *Review of Educational Research*, volume 66, pages 39–51, 1996.

[24] A. Chalamandaris, S. Raptis, P. Tsiakoulis, and S. Karabetsos. Enhancing accessibility of web content for the print-impaired and blind people. In *Proceedings of the 5th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society on HCI and Usability for e-Inclusion*, USAB '09, pages 249–263, Berlin, Heidelberg, 2009. Springer-Verlag.

[25] C.-H. Chang, M. Kayed, R. Girgis, and K. Shaalan. A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on*, 18(10):1411 –1428, oct. 2006.

[26] C.-S. Chen, S.-F. Chang, and C.-H. Liu. Understanding knowledge-sharing motivation, incentive mechanisms, and satisfaction in virtual communities. *Social Behavior and Personality: an international journal*, 40(4):639–647, May 2012.

[27] P. P.-S. Chen. English sentence structure and entity-relationship diagrams. *Information Sciences*, 29(2-3):127 – 149, 1983.

[28] C. Cooley. *Human Nature & Social Order - Ppr*. Social Science Classics Series. Transaction Pub, 1964.

[29] S. Crawford. How gamification works. http://electronics.howstuffworks.com/gamification5.htm. Last visited September 2012.

[30] J. W. Creswell. Chapter 1: A framework for design, in research design: qualitative, quantitative and mixed methods. Sage Publications.

[31] M. Debevc, P. Kosec, and A. Holzinger. Improving multimodal web accessibility for deaf people: sign language interpreter module. *Multimedia Tools and Applications*, 54:181–199, 2011.

[32] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI'05, pages 1034–1041, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[33] D. Downey, S. Schoenmackers, and O. Etzioni. Sparse information extraction: Unsupervised language models to the rescue. In *In Proc. of ACL*, pages 696–703, 2007.

[34] W. Du and M. J. Atallah. Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on New security paradigms*, NSPW '01, pages 13–22, New York, NY, USA, 2001. ACM.

[35] C. Dugan, M. Muller, D. R. Millen, W. Geyer, B. Brownholtz, and M. Moore. The dogear game: a social bookmark recommender system. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 387–390, New York, NY, USA, 2007. ACM.

[36] T. B. D. Florian Boudinet, Julien Allali. Report on alipi browser plugin. https://fboudinet.frenchdev.com/files/report-fboudinet-i3-servelots.pdf. Last visited March 2013.

[37] D. Freitag and A. Mccallum. Information extraction with hmm structures learned by stochastic optimization. In *In Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 584–589. AAAI Press, 2000.

[38] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys.*, 1(3-4):219–234, Dec. 2003.

[39] R. Ghosh and M. Dekhil. Discovering user profiles. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1233–1234, New York, NY, USA, 2009. ACM.

[40] M. Golemati, C. Halatsis, C. Vassilakis, A. Katifori, and Peloponnese. A context-based adaptive visualization environment. In *Proceedings of the conference on Information Visualization*, IV '06, pages 62–67, Washington, DC, USA, 2006. IEEE Computer Society.

[41] M. Golemati, A. Katifori, C. Vassilakis, G. Lepouras, and C. Halatsis. Creating an ontology for the user profile: Method and applications. In *In Proceedings of the First International Conference on Research Challenges in Information Science (RCIS)*, 2007.

[42] E. Greengrass. Information retrieval: A survey, 2000.

[43] A. P. P. Gregory Gutin. *The Traveling Salesman Problem and Its Variations*. Springer, 2006.

[44] D. A. Grier. *When Computers were Human*. Princeton University Press, 2005.

[45] T. Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):4 – 13, 2008.

[46] Z. GuoDong, S. Jian, Z. Jie, and Z. Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[47] T. Hasegawa, S. Sekine, and R. Grishman. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[48] S. Hatch, D. Thomsen, and J. Waldron. Extrinsic rewards and motivation. https://www.appliedsportpsych.org/resource-center/coaches/articles/extrinsicrewards. Last visited September 2012.

[49] S. L. Henry. Understanding web accessibility. http://uiaccess.com/understanding.html. Last visited Jan. 2013.

[50] B. A. Huberman, D. M. Romero, and F. Wu. Crowdsourcing, attention and productivity. *J. Inf. Sci.*, 35(6):758–765, Dec. 2009.

[51] P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on Human Computation (HCOMP 2010)*, pages 64–67. ACM, 2010.

[52] M. Ito, S. Baumer, M. Bittanti, d. boyd, R. Cody, B. Herr-Stephenson, H. Horst, P. Lange, D. Mahendran, K. Martinez, et al. *Hanging Out, Messing Around, and Geeking Out: Kids Living and Learning with New Media*. John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. MIT Press, 2009.

[53] Janastu. Alipi code repository. http://janastu.org/technoscience/index.php/AlipiGitHub. Last visited March 2012.

[54] T. S. Jayram, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Zhu. Avatar information extraction system. *IEEE Data Eng. Bull.*, pages 40–48, 2006.

[55] N. Kaklanis, K. Votis, K. Moustakas, and D. Tzovaras. 3d hapticwebbrowser: towards universal web navigation for the visually impaired. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, W4A '10, pages 25:1–25:2, New York, NY, USA, 2010. ACM.

[56] N. Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, ACLdemo '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[57] N. Kaufmann and T. Schulze. Worker motivation in crowdsourcing and human computation. In *AAAI workshop on human computation (HCOMP), 8 August, San Francisco, USA*. AAAI Press, 2011.

[58] B. Kelly, S. Lewthwaite, and D. Sloan. Developing countries; developing experiences: approaches to accessibility for the real world. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, W4A '10, pages 3:1–3:4, New York, NY, USA, 2010. ACM.

[59] B. Kelly, D. Sloan, L. Phipps, H. Petrie, and F. Hamilton. Forcing standardization or accommodating diversity?: a framework for applying the wcag in the real world. In *Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility (W4A)*, W4A '05, pages 46–54, New York, NY, USA, 2005. ACM.

[60] S. Khanna, A. Ratan, J. Davis, and W. Thies. Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 12:1–12:10, New York, NY, USA, 2010. ACM.

[61] R. A. Khot and K. Srinathan. Gofish: fishing thousand words worth a picture. In *Proceedings of the 2010 international conference on Interaction Design & International Development*, IHCI'10, pages 73–82, Swinton, UK, UK, 2010. British Computer Society.

[62] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[63] M. Kliehm. Accessible Web 2.0 Applications with WAI-ARIA. http://www.alistapart.com/articles/waiaria, Apr. 2007. Last visited Jan. 2013.

[64] A. J. Ko and P. K. Chilana. How power users help and hinder open bug reporting. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1665–1674, New York, NY, USA, 2010. ACM.

[65] R. Koster. *A theory of fun for game design*. Paraglyph Series. Paraglyph Press, 2005.

[66] Y.-l. Kuo, J.-C. Lee, K.-y. Chiang, R. Wang, E. Shen, C.-w. Chan, and J. Y.-j. Hsu. Community-based game design: experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 15–22, New York, NY, USA, 2009. ACM.

[67] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[68] E. Law and L. Von Ahn. *Human Computation*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2011.

[69] E. L. M. Law, L. von Ahn, R. B. Dannenberg, and M. Crawford. Tagatune: A game for music and sound annotation. In *ISMIR'07*, pages 361–364, 2007.

[70] B. Lin and P. A. Dinda. Experiences with scheduling and mapping games for adaptive distributed systems: summary. In *Proceedings of the 6th international conference on Autonomic computing*, ICAC '09, pages 73–74, New York, NY, USA, 2009. ACM.

[71] L. Lyons. Internet usage by people different age group. http://www.gallup.com/poll/10993/Internet-Use-Whats-Age-Got-It.aspx. Last visited Feb. 2013.

[72] D. S. J. M. R. Garey. *Computers and Intractibility: A guide to the theory of NP-completeness*. W. H. Freeman, 1979.

[73] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 77–85, New York, NY, USA, 2009. ACM.

[74] R. McDonald, K. Crammer, and F. Pereira. Flexible text segmentation with structured multilabel classification. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 987–994, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[75] P. McNamee, H. T. Dang, H. Simpson, P. Schone, and S. Strassel. An evaluation of technologies for knowledge base population. In *LREC'10*, pages –1–1, 2010.

[76] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

[77] S. Mirri, P. Salomoni, and C. Prandi. Augment browsing and standard profiling for enhancing web accessibility. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, W4A '11, pages 5:1–5:10, New York, NY, USA, 2011. ACM.

[78] Mozilla. Universal subtitles - transcribe, caption, translate, and subtitle videos. http://www.universalsubtitles.org/en/, 2011.

[79] K. K. Nam, M. S. Ackerman, and L. A. Adamic. Questions in, knowledge in?: a study of naver's question answering community. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, pages 779–788, New York, NY, USA, 2009. ACM.

[80] W. L. Neuman. Social research methods. Pearson Education. 6th Edition.

[81] S. Noor and K. Martinez. Using social data as context for making recommendations: an ontology based approach. In *Proceedings of the 1st Workshop on Context, Information and Ontologies*, CIAO '09, pages 7:1–7:8, New York, NY, USA, 2009. ACM.

[82] A. J. Quinn and B. B. Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1403–1412, New York, NY, USA, 2011. ACM.

[83] S. Reiss. Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of General Psychology*, 8(3):179–193, 2004.

[84] M. Rowe and F. Ciravegna. Getting to me exporting semantic social network information from facebook, 2008.

[85] S. Sarawagi. Information extraction. *Found. Trends databases*, 1(3):261–377, Mar. 2008.

[86] J. F. Saray Villamizar, B. Encelle, Y. Prié, and P.-A. Champin. An adaptive videos enrichment system based on decision trees for people with sensory disabilities. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, W4A '11, pages 7:1–7:4, New York, NY, USA, 2011. ACM.

[87] S. Sekine. Named Entity: History and Future. 2004.

[88] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In P. J. Hinds, J. C. Tang, J. Wang, J. E. Bardram, and N. Ducheneaut, editors, *CSCW*, pages 275–284. ACM, 2011.

[89] K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. *Intelligent Systems, IEEE*, 23(3):50 –60, may-june 2008.

[90] M. Sutterer, O. Droegehorn, and K. David. Upos: User profile ontology with situation-dependent preferences support. In *Advances in Computer-Human Interaction, 2008 First International Conference on*, pages 230 –235, feb. 2008.

[91] P. Sweetser and P. Wyeth. Gameflow: a model for evaluating player enjoyment in games. *Comput. Entertain.*, 3(3):3–3, July 2005.

[92] Z. Syed, E. Viegas, and S. Parastatidis. Automatic discovery of semantic relations using mindnet. *LREC*, 2010.

[93] U. T. Book review: Intrinsic motivation, extrinsic rewards, and divergent views of reality. *Educational Psychology Review*, 15(3):311–325, 2003.

[94] H. Takagi, S. Kawanaka, M. Kobayashi, T. Itoh, and C. Asakawa. Social accessibility: achieving accessibility through collaborative metadata authoring. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, Assets '08, pages 193–200, New York, NY, USA, 2008. ACM.

[95] TED. Ted translations. http://www.ted.com/OpenTranslationProject, 2011.

[96] V. Topac and V. Stoicu-Tivadar. Patient empowerment by increasing information accessibility in a telecare system. *Stud Health Technol Inform*, 169, 2011.

[97] I. Tuomi. Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *J. Manage. Inf. Syst.*, 16(3):103–117, Dec. 1999.

[98] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[99] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, Aug. 2008.

[100] L. von Ahn, S. Ginosar, M. Kedia, and M. Blu. Improving image search with phetch. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2007.

[101] L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pages 75–78, New York, NY, USA, 2006. ACM.

[102] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pages 55–64, New York, NY, USA, 2006. ACM.

[103] W3C. Web accessibility initiative (wai). http://www.w3.org/WAI/, 2011. Last visited March 2012.

[104] W. M. Watanabe. Facilita: reading assistance to the functionally illiterate. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, W4A '10, pages 7:1–7:2, New York, NY, USA, 2010. ACM.

[105] W. M. Watanabe, A. Candido, Jr., M. A. Amâncio, M. de Oliveira, T. A. S. Pardo, R. P. M. Fortes, and S. M. Aluísio. Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, W4A '10, pages 8:1–8:9, New York, NY, USA, 2010. ACM.

[106] S. C. Weller. Cultural consensus theory: Applications and frequently asked questions. *Field Methods*, 19(4):339–368, 2007.

[107] M. Wick, A. Culotta, and A. McCallum. Learning field compatibilities to extract database records from unstructured text. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language*

Processing, EMNLP '06, pages 603–611, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[108] F. Wu, D. M. Wilkinson, and B. A. Huberman. Feedback loops of attention in peer production. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, pages 409–415, Washington, DC, USA, 2009. IEEE Computer Society.

[109] M.-C. Yuen, I. King, and K.-S. Leung. A survey of crowdsourcing systems. In *Privacy, security, risk and trust (passat), 2011 IEEE Third International Conference on Social Computing (socialcom)*, pages 766 –773, oct. 2011.

[110] J. Zimmerman, J. Forlizzi, and S. Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 493–502, New York, NY, USA, 2007. ACM.