

Aalto University
School of Science
Master's Programme in ICT Innovation

Rowshan Sathi

Personal Information Center(PIC)- A Data Integration Service for the Private Cloud

Master's Thesis

Berlin, July 5 , 2015

Supervisor: Professor Jukka K. Nurminen, Aalto University
Professor Dr.Axel Küpper, Technical University Berlin

Instructor: Dr.Gökhan Coskun

Aalto University School of Science Degree Programme in Computer Science and Engineering Master's Programme in ICT Innovation		ABSTRACT OF THE MASTER'S THESIS	
Author: Rowshan Jahan Sathi			
Title: Personal Information Center(PIC)- A Data Integration Service for the Private Cloud			
Number of pages: 74	Date: 05.07.2015	Language: English	
Professorship:		Code:	
Supervisor: Professor Jukka Nurminen, Aalto University Professor Dr.Axel Küpper, Technical University Berlin			
Advisor: Dr.Gökhan Coskun			
Abstract: <p>Managing information has become an extra load for our everyday life but creating a personal information center (PIC) solves this problem easily. A PIC makes it easier to see content such as email messages, weather information, news items, and even information from local storage by using one view or user interface. Additionally, it fulfills the real world requirements like accessibility, around-the-clock availability of service, and it solves the device constraints problem by creating a common presentation format for all kinds of devices. An importer works spontaneously as a fetcher, parser and processor to process the information to create the nodes content. An analyzer works as an extractor and creates vocabulary terms automatically to autotag the node content. Finally, a filtering system works for finding similar nodes to create one common presentation with all matched contents. All the works in the PIC is done automatically without any users' interaction, only the source of information is defined by the users. The analyzer uses machine learning Naive Bayes approach to extract keyphrases from the contents. The PIC uses an advanced filtering system to find similarity between nodes and to create a common presentation for all devices.</p>			
Keywords:			

Acknowledgements

I would like to express my sincere gratitude and appreciation to my supervisor Dr.Goekhan Coskun (TU Berlin) and Steffen Druessedow (Deutsche Telekom AG) for their guidance, advice, encouragement and insight throughout the master thesis period.

I would like to express my special thanks to Prof.Dr.Axel Küpper for extending my thesis time period and giving me a chance to finish my master thesis work.

Finally, I would like to thank EIT ICT Labs Master School for giving me a chance to attend this master program and all the people in Seamless Network Control group at Deutsche Telekom AG.

Contents

1. Introduction	1
1.1. Thesis Motivations and Objectives	2
1.2. Research Questions	2
1.3. Approaches	3
1.4. Thesis outline	3
2. Background	5
2.1. Data Aggregation	5
2.2. Natural Language Processing (NLP)	5
2.2.1. Information retrieval	5
2.2.2. Automatic Keyphrase extraction	6
2.2.3. Branches of keyword extraction	6
2.2.4. Features of the keyword extraction	6
2.3. Text Analysis	7
2.3.1. Benefits of text analysis	7
2.4. Information Filtering	7
2.4.1. Content based filtering	7
2.4.2. Challenges of collaborative filtering	8
2.5. Content management system(CMS)	8
2.5.1. Drupal	8
2.5.2. Modules	8
2.6. knowledge based system	9
2.7. Agent Based Technology	9
2.8. Semantic Web Based Technology	9
2.9. RDF	9
2.10. Triples	9
2.11. Performance measure	9
2.11.1. Precisions	9
2.11.2. Recall	9
2.11.3. F-measure	10
3. Related Works	11
3.1. Agent based approach	11
3.2. Neural Network based approaches	13

3.3. Machine learning Approaches	13
3.4. Co-occurrence	14
3.5. Semantic web based approaches	15
4. Problem Analysis	17
4.1. What is personal information center?	17
4.2. Collecting data	18
4.3. Data categorization	18
4.4. Aggregate data from multiple sources	18
4.5. Filtering data	19
5. Decisions	21
6. System Design	23
6.1. Content creation layer	23
6.2. Extraction layer	24
6.3. Filtering Layer	26
6.4. Display Layer	27
7. Deployment	31
8. Implementation	33
8.1. Content creation	35
8.2. Collecting information	35
8.3. Keyphrase extraction	36
8.4. Filtering	44
8.5. Presentation	45
8.6. Protocols used for this implementation	45
8.7. Modules and features	45
8.8. Final display	45
8.9. Challenges for Implementation	46
9. Evaluation	49
9.1. Performance of the hardware	49
9.2. Data collection using our designed system	49
9.3. PIC a system that aggregate data	49
9.4. Filtering system for personalized web content	50
9.5. Presentation of PIC	50
9.6. Data privacy of PIC	50
9.7. Drupal CMS for PIC	50
10. Discussion	53
11. Conclusion and Future Work	55
11.1. Conclusion	55
11.2. Future Work	56

12. List of Tables	57
13. List of Figures	59
14. Listings	61
15. Bibliography	63
A. Evaluation Matrix	65

1 Introduction

The internet is an open source of information with different categories of data such as weather, entertainment, politics, sports, and live web cam pictures. Therefore it is very difficult and confusing to decide which type of information to read and which one is the most important and reliable according to our needs. On the other hand, people from different profession such as people like university professor and lecturer, IT professionals, bankers and researchers etc do not have enough time to search different news items separately as well as to check all their email messages from different email accounts. For example, it is important for an IT professional to know information about new and innovative software technology and hardware devices but the internet is a huge source of information and to search for their preferred information takes up valuable time from their day. Another scenario is if a researcher wants to see his important emails related to one meeting in Paris and wants to check his other meeting schedule on that day in his personal web calendar, then it is a great hassle because then the researcher needs to do it manually.

Park et al.2003 mentioned that normally people visits a limited number of sites according to their needs and interests. Usually, they save the sites in their own web browser favorites so whenever they want they can visit those sites and get updated information according to their needs [1]. Thus, building one service for the users would be one solution to these problems. The main benefits of this service would be around-the-clock accessibility and contents would customizable according to the users preferences. In other words, building a single personal information center (PIC) is a solution for this. In this thesis PIC is a private cloud service to aggregate data from multiple sources which can solve the need of creating a personalized content filtering system.

The number of internet users is increasing day by day. From 1995 to till March 2014 the total number of population using internet is 2,937 million¹ which is 40.7% of the world population, whereas in 2012 it was 35.5% of the world population. This increasing number of internet users' also opens a new market for the software industry to develop a system that can fulfill the information needs of this huge number of users according to their interests in a personalized way.

Creating this kind of service or platform requires to have an appropriate system design which is not an easy task. Rossi et al.2001 mentioned that it is very challenging to create personalized web due to the involvement of lots of technologies, therefore, they emphasized on design view in their paper. They analyzed some of the existing scenarios of personalized web content to design their personalized web [2] but to only design a service with good design

¹Source: <http://www.internetworldstats.com/emarketing.htm>

model does not make it acceptable to the user. There are already many framework such as PIA [3], PI-Agent [4], PEA [5], Fab[6], Letizia [7], OWLIR [8], PIRATES [9] which have been developed to fulfill this need. Although all these framework used advanced technology such as agent based, neural network based etc they were not widely accepted by people as they could not satisfy the real world requirements like- on demand availability or scalability, and some of them cannot support devices like mobile, PDA, pictureframe etc. There are also privacy issues to consider when accessing users sensitive or personal data. Therefore, it has become important to develop one service that will fulfill all these real world requirements so that user can use it from any device from everywhere at any time and will protect users' sensitive data.

Hence, in order to find out the basic requirements of the system design for such kind of system and to implement them, a root cause analysis of the problem is an important aspect of this research study.

1.1. Thesis Motivations and Objectives

As per the previous discussion of this chapter, we can say that the main idea behind this research is to provide users a seamless service by creating a personal information center without any device constraints and with one common presentation format for all devices. This service will be, scalable and easily accessible by the users whilst maintaining their data security and privacy. After considering all the needs and issues for creating such a service we can progress to the main goal of our research. The main goal of this thesis is to develop a service that aggregates information from multiple sources. These sources can be from the web or from personal storage or camera in the format of an image or file. In addition, the proposed service will be deployed under private cloud platform to ensure the availability of the service on premises and off premises or from wherever the user wishes to access its.

1.2. Research Questions

As PIC is dealing with data from private and public sources it is a great challenge to aggregate these data. The proposed study is going to examine the following questions:

- How to collect data from multiples sources ?
- How to build a service to aggregate collected data from multiple sources?
- How to analyze the aggregated data and find similarity among them automatically inside the designed system or service ?
- Finally, how to solve the presentation related issues for the devices with different platforms?

1.3. Approaches

In this thesis the whole process of data aggregation consists of following steps -

- data collection from multiples sources by developing a system.
- machine learning approach based tools to build model for the purposes of the keyword extraction.
- filtering algorithm to find the similar content and finally, creating a common view for

multi-device platforms.

1.4. Thesis outline

This paper is structured as follows: **Chapter 2** presents background of the proposed thesis. **Chapter 3** presents the relevant research work from the researchers. **Chapter 4** presents the problems analysis of the proposed research study. **Chapter 5** Decisions. **Chapter 6** describes the different steps of the proposed research design. **Chapter 7** describes deployment of the system design. **Chapter 8** describes the implementation of the proposed research work in details. **Chapter 9** presents the evaluation and prototyping of the proposed research implementation. **Chapter 10** presents a discussion of the research according to the relevant research. **Finally**, chapter 11 concludes the overall talk of this research by presenting a direction and guidelines for the future.

2 Background

This chapter chronologically presents some basic knowledge about the terms, technologies and approaches used throughout this thesis.

2.1. Data Aggregation

Data aggregation is a process of collecting data or information from different sources for a specific purpose such as- content analysis or statistical analysis. We can analyze the content or information by using Natural Language Processing(NLP).

2.2. Natural Language Processing (NLP)

Natural language processing is a branch of computer science focused on the interactions between computer and human language by developing system. It also is known as computational linguistics. NLP solves many real world problems such as face recognition. The major tasks of NLP is Automatic summarization, Coreference resolution, Natural language generation, Information retrieval, Information filtering, parts of speech tagging (POS), parsing, stemming, text categorization etc. In the following sections we are going to introduce information retrieval and information filtering and it related terms that used in this research.

2.2.1. Information retrieval

Information retrieval is a technique for storing, searching and retrieving information according to an information need from a collect of information resources. In other words, retrieval of unstructured or text based information is referred to as information retrieval. This text of IR system composed of documents and terms. Documents is journal paper, book, articles, e-mail messages, source code, web pages etc . Whereas, terms are the word or phrases in a document. Though it is a branch of computer science but it uses some of the NLP methods such as- n-grams extraction, stemming etc to process the information or text. There are many ways to retrieve information. Automated information retrieval is easy to use and it reduces the information overload.

2.2.2. Automatic Keyphrase extraction

Automatic keyphrase extraction is a part of information retrieval. Automatic keyphrase extraction is extracting important keywords from a document. Turney(1999) [10] preferred to call it keyphrases rather calling it keywords and as a reason he mentioned that most of the keywords

from any document consists of phrases with two or more words. He define the automatic keyphrase extraction as- the automatic selection of important, topical phrases from within the body of a document. Automatic keyphrase extraction is a special case of the more general task of automatic keyphrase generation, in which the generated phrases do not necessarily appear in the body of the given document. We can extract keyphrase using free indexing and with a controlled vocabulary known as keyword assignment.

2.2.3. Branches of keyword extraction

- Free indexing: In free indexing, keyphrases are extracted from a document without using a vocabulary but on the basis of features like frequency and length.
- Keyword assignment: In keyword assignment, keyphrases are chosen from a controlled vocabulary of terms.

2.2.4. Features of the keyword extraction

There are some features that almost all of the keyword indexing method are using-

1. Term frequency
2. TF*IDF
3. First Occurrence
4. Keyphraseness
5. Phrase length
6. Node spread
7. Spread
8. Semantic relatedness
9. Wikipedia-based keyphraseness
10. Inverse Wikipedia linkage

In this section we are going to talk about Term frequency and TF*IDF as it is one of the most widely used algorithm related to this research field.

- Term frequency: Find out the mostly occurred words or keyphrases in a particular document. More specifically, $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d .
- TF*IDF: The most frequent terms in a document is term frequency. Whereas, Inverse Document Frequency is rarely used terms in a document. We calculate it as follows-

$$idf_i = \log(|D| / |\{j : t_i \in d_j\}|)$$

, where $|D|$ is total number of document in the corpus. $|\{j : t_i \in d_j\}|$ is the number of document the term t_i appears. Therefore, we can say that TF*IDF combines term frequency and inverse document frequency. TF*IDF is high if a keyphrase appears rarely in

a document therefore, the keyphrase is more significant.

2.3. Text Analysis

Text analysis or content analysis is the processing of a page content or document text using algorithm or machine learning approaches. When we analyze the text document or page content we think about the following questions:

- What are the types of the data?
- What are the source that data is coming from?
- How to analyze text? What would be the algorithm, approaches etc?
- What would be the criteria of the text analysis?
- How the see the analyze content?

2.3.1. Benefits of text analysis

The main benefits of the text analysis can be to build a tagging system with analyzed text. Clustering different text document according to the analyzed word.

2.4. Information Filtering

Information filtering is a process of extracting information or delivering information from different sources that the users is likely to find interesting and useful. In information filtering users only see the data that is extracted. It is involved with large amount of data. It is most often known as recommender system. There are two major types of recommender system exists. Content based and collaborative filtering.

2.4.1. Content based filtering

Selects an item based on the correlation between the content of the item and the users preferences. Vector space model, latent semantic indexing and relevance feedback is content based filtering. There are several issues to consider when implementing content based filtering. They are-

- First, terms can be assigned automatically or manually.
- Second, the matching between the user profiles and the items has to be done in an meaningful way.
- Finally, a learning algorithm is able to learn the users profile based on the seen item and this learning algorithm can give feedback based on the user profile.

Collaborative filtering system Selects items based on the correlation between users with similar preferences. There are several types of collaborative filtering technique such as- memory based, Model based and hybrid. There are two types of collaborative filtering. they are-

1. User based. For this type of filtering system nearest neighbor algorithm is used.
2. Item based. It follows an item-centric manner to filter information.

2.4.2. Challenges of collaborative filtering

The main challenges of collaborative filtering is -

- data sparsity
- scalability
- synonyms
- Grey sheep
- Shilling attacks
- Diversity and the Long Tail

There is also hybrid filtering technique that merges both of the approaches together. Other filtering technologies are- demographic filtering, economic filtering etc.

2.5. Content management system(CMS)

CMS is a application to store, publish and arrange web content in it. There are many CMS such as-drupal, typo3, wordpress etc. The main benefits of using CMS is that user does not need to code. It has some features, library, plugins or modules that can help to build web pages.

2.5.1. Drupal

Drupal is an open source content management system. It is actively developed,maintain and updated by the drupal community. The technology stack of drupal is- a server, operating system, database, webserver, php.Drupal is a modular system and it is written in php. It different module for different purposes. Drupal modules are also extendible. It is possible to modify and create a new module from the scratch.

2.5.2. Modules

Modules are consists of lots of functions and methods. Drupal has core module and contributed modules. Contributed module offers extra features such as custom content type and content listing. The most commonly used contributed drupal modules are-

- Views
- Content Construction Kit (CCK)
- Rules
- Features
- Context etc

2.6. knowledge based system

Knowledge based system is a software system that is used to solve complex problem with knowledge base reasoning. The term knowledge base is used to store complex unstructured information from different sources. Usually, it is an intelligent system.

2.7. Agent Based Technology

It's an intelligent system where all the problem will be monitored and solved by specific agents. Agents are the computer programs consists of advanced algorithms, methods or functions to solve problems in an intelligent manner.

2.8. Semantic Web Based Technology

Semantic web is also called web of data¹.According to w3c,"The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries" [11].

2.9. RDF

In full form resource description framework. It is a framework to create triples. It presents resources as a graph form.

2.10. Triples

A triple consists of three parts: a subject, a predicate and an object.

2.11. Performance measure

The measure we are going to use in the evaluation is F-measure, precisions and recall.

2.11.1. Precisions

Precisions measures the number of related document among all the retrieved documents.

2.11.2. Recall

Recall is the proportion of number of retrieved document to the number of total related document.

2.11.3. F-measure

F-measure is a combination of precision and recall. the origin formula to calculate

$$F - measure = 2 * (precision \cdot recall) / (precision + recall)$$

¹<http://www.w3.org/2001/sw/>

3 Related Works

This chapter reviews the literature related to this thesis. In this literature review we talked about many approaches that can give an idea about past research work conducted by many researchers. It can also show a direction to design and implement PIC from the past research experience. Therefore, it is very important to review previous study and research to create a service like PIC and also to gain users acceptability. The next section of this research discusses various approaches; agent based approach, neural network based approach, machine learning approach, semantic web based approach and some other approaches like co-occurrence.

3.1. Agent based approach

1. PIA [3] Personal information agent is an agent based approach[12][13] for information filtering where an agent delivers information by accessing, filtering and integrating information in its own system environment. In this study, PIA permits the accessibility of the information by www ,email, SMS,MMS and J2ME clients. In this system a user give an input request to find his preferred information according to the users choice or even a user can select input from a predefined keyword lists created by using collaborative filtering. This system extracts keywords from a document using TF*IDF [14] algorithm and PIA has a keyword assistant to extract keywords from a document. It has four types of agent and in addition, one agent for each user. Information extracting agent extract information, whereas, there is another agent for filtering. The Figure-3.1 presents the system design of PIA system. Agents build model using different algorithms. More precisely,PIA has different agent to perform different task. It uses TF*IDF algorithm to extract keyphrases from the collected information set and this algorithm TF*IDF is the main concern for the efficient retrieval. PIA fulfills mostly all of the real world requirements like availability, scalability and adaptation to any kind of mobile device, PDA.
2. **P-Tango** [15] grabs the news articles with the users' highest interest and shows in a personalized common newspaper view. The filtering method it is using is a combination of collaborative and content based filtering. But it can support any types of filtering algorithms of technique.
3. Fab [6] used for information retrieval. It uses content based filtering to create users profile but later it finds similar users for collaborative recommendation. Though it uses a combination of two recommendation system but mainly to generate an effective and functional

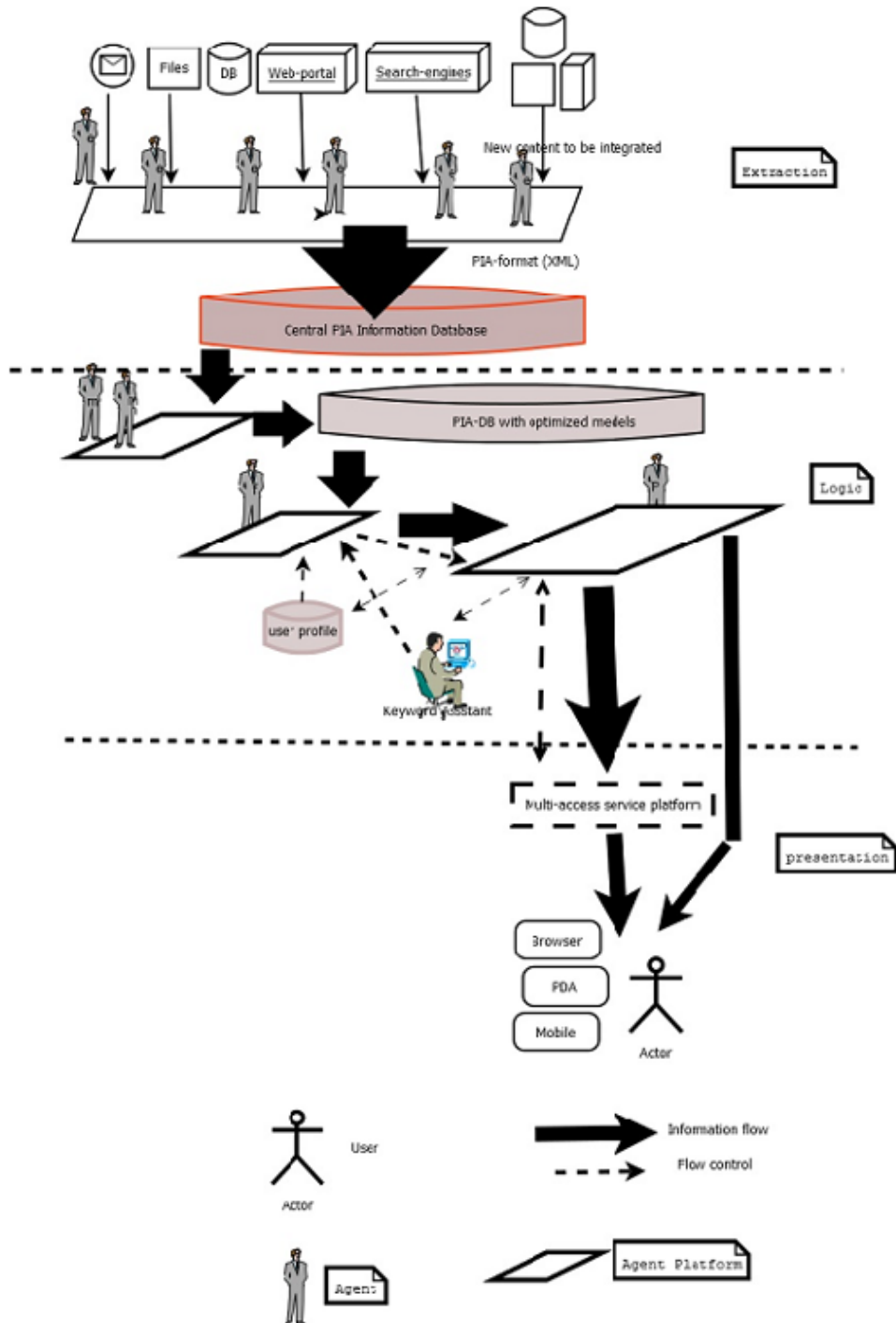


Figure 3.1.: PIA system design[3]

profile it always uses content based filtering.

4. As we are receiving lots of email everyday and sometimes it happens that we do not have enough time to read them and sometimes it is also impossible to select which one is important email. In order to get rid of this problem PEA [5] is built. This framework saves our valuable time by filtering important emails. It builds a user model and collect users interests and behaviors to predict the future action using an evolutionary algorithm. This algorithm finds out the user models that are more close to the present information needs. Finally, It ranks the emails according to their relevance and presents it to the users.
5. Letizia [7] is an intelligent system for web browsing. It is built to find out who is browsing interesting information in the internet. Usually, it analyzes users behavior by following hyperlinks in an html document. It tracks user behavior and uses various heuristics technique to predict potential relevant documents for the user without using any retrieval tools. It also suggests which document should be ignored to the user. However it is a bit time consuming as learning about a person takes times and there are many rules and parameters associated with its learning process.
6. Amalthea [16] is a multiagent based personalized system. It tries to collect relevant information about a user and these information are collected from different sources. For information filtering and information discovery Amalthea has two agents and these agents use evolution programming to perform several task to filter and discover information.

3.2. Neural Network based approaches

1. In this paper interested users get articles from different sources automatically using PIAgent [4]. It uses neural network based approaches to separate similar articles. As PIAgent uses neural network approaches to generate a training document set it becomes a little expensive, but these neural network based approaches make the text classification more accurate.
2. In paper [17] a neural network based reinforcement learning method is used to find the current state of a user. This system learns about the user profile by observing his interacting behavior in the system and estimates the users relevant feedback to a document. To collect this information WAIR has developed three agents -a user-interface agent, a Web-document retrieval agent, and a learning agent. A user-interface agent observes the users behavior and learns his web browsing behaviors from that. A web-document retrieval agent collects the users interests from user-interface agent and retrieves a candidate document set. Here a candidate document set is an html document. Finally, the learning agent learns or adapts the state of the user profile by using reinforcement learning method. A user agent sends data to the learning agent.

3.3. Machine learning Approaches

1. KEA[18] uses Naive Bayes machine learning classifier to build a model. Its keyword extraction algorithm has two parts- training and keyphrase extraction. It is the state of the

art for keyphrases extraction. The Figure-3.2 shows the keyphrase extraction using KEA algorithm. Precision for keyphrase extraction from generic web pages is about 18% according to KEA team report. A domain specific vocabularies improves the results up to 28.3% for recall and 26.1% for precision [19] than keyword extraction without a vocabulary.

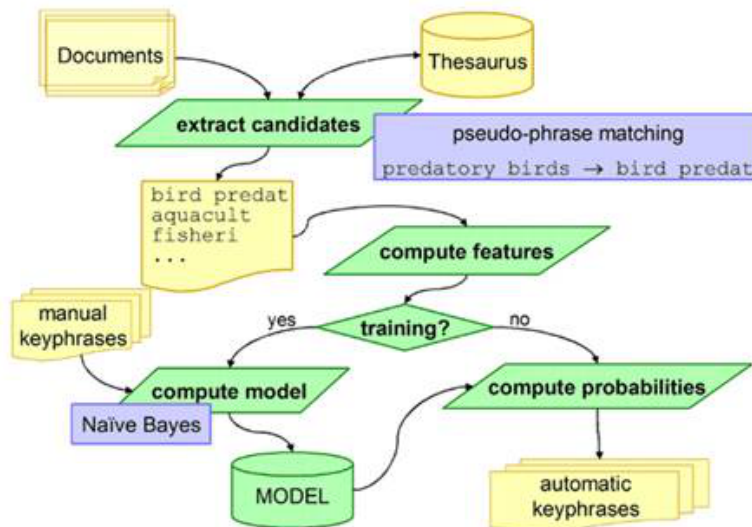


Figure 3.2.: Keyphrase Extraction using KEA[19]

- In [20] uses GenEx to implement an approach suggested by Turney. GenEx algorithm is a combination of genetic algorithm and heuristic rules. But to extract keyword it takes time. It provides precision and recall value nearly same as KEA.

3.4. Co-occurrence

- In [21] a new algorithm is proposed to extract keyword from a single document. In this study the authors extract the frequent term and then count the co-occurrences of a term and frequent term. They used x_2 measure to find the frequent term of a co-occurrence in their paper.
- In paper[22] extracts keyword automatically using statistical approach. Candidate phrases extraction is completed in a combination of graph methods(TaxtRank¹) and statistical methods (TF*IDF). The algorithm is divided in the three parts-text processing, keyword extraction, keyphrase extraction. For keyphrase extraction they mentioned NLP method where n-grams are extracted. keyword and keyphrases are ordered by TF*IDF score.

¹ <https://metacpan.org/pod/Text::Categorize::Textrank>

3. In [23] a new feature is proposed based on document frequency and statistical values. This study uses a similarity measure to calculate similarity of different document. The authors mentioned three methods to categorize text and using these methods provides higher the performance than [24] for text categorization.

3.5. Semantic web based approaches

1. OWLIR [8] is a semantic retrieval system based on the semantic indexing method and it has three module to use this semantic indexing approach. They extract keyword from free text using AeroText² system and keyphrases are considered as RDF triples. But to obtain an RDF triples, an extracted keyphrase has to be enriched with reasoning and semantic rules. Once these triples are indexed the details of the indexing and ranking is deleted.
2. This paper [9] represents a new auto tagging approach where domain ontology is used to build a tagging tools. Along with ontology domain they use tags, keyphrase extraction method etc for generating auto tag. The framework they build is PIRATES (Personalized Intelligent tag Recommender and Annotator TESTbed)

Therefore, the literature review shows that there are a lot of framework exists for personalized web content management. But none of them fulfills all kinds of realworld needs. From the literature reviews, we can see that PIA system is an advanced framework that almost fulfill all of the realworld requirements. But extraction process is not very consistent and robust. on the other hand, KEA is an algorithm that uses machine learning approach to extract keyphrases from a document which is very consistent and robust for keyphrase extraction. Therefore, we try to build a frame work that follows PIA system framework but internal keyword extraction system and filtering system is using more advanced algorithms and approaches.

²<http://www.rocketsoftware.com/products/rocket-aerotext>

4 Problem Analysis

In this research we are trying to develop a service that can aggregate data from multiple sources. This service will be presentable for any kind of devices and available for the users from wherever the users wishes to access. Therefore, in this chapter we are going to analyze the following questions in details in order to solve the research problems for such kind of service development:

- What is personal information center?
- How to collect data and categorize them?
- How to aggregate collected data from multiple sources?
- Why and how to analyze the aggregated data ?
- How to find the similarity of different types of aggregated content?

4.1. What is personal information center?

Personal Information center (PIC) is a data aggregation service for the private cloud. PIC aggregates data from the different sources according to the user's preference. The data sources are defined according to the users and the defined types of the data is the news and RSS feeds, weather information, camera pictures, calendar events as well as personal emails. Users can see the web content according to their preference. Even they are able to customize their personal information center, for example they can delete, update and edit their desired content in the personal information center.

The question is what are the main types of the data sources we want to aggregate? From our analysis we found two main types of the data sources in the web and they are- private and public data source. Private data are the most sensitive data such as email, calendar events data and information from personal storage (such as important document or image etc) or personal camera image. On the other hand, public data are rss feeds, news feeds and weather information. Personal information center is not concerned about the origin of data sources, for example whether the data source is from CNN or BCC. Rather it is more concerned about the privacy of data because some of the data is coming from the users personal data such as emails, calendar and personal storage data.

4.2. Collecting data

We have mentioned earlier that PIC is not concerned about the data sources and we want to collect data as rss feeds, emails, calendar events, camera image etc. To do this we need to build

a system that can collect or grab these data from multiple sources. It can be an information retrieval system, a query editor or an automated search engine or platform that can collect information from different sources. Thus, we can say that this is an important challenge for our personal information center (PIC).

4.3. Data categorization

In PIC data are coming from multiples sources as public and private data. Therefore, to categorize the data is important. It is also important to define sensitive data in one category. For example- News data under News items. Weather data under weather items, emails are under email items and camera image will also be defined under another item.

4.4. Aggregate data from multiple sources

Data aggregation means-collecting data and then processing data for a specific purpose such as statistical analysis or content analysis. This study mainly focuses on the content analysis. Why do we need to analyze the content? The main purpose of content analysis is to find the most important keyphrases in the document or most commonly used words. After analyzing the content we extract the keyphrases from that document. Therefore, to find a most important word in a document we need to follow some rules or algorithm or approaches that helps to determine the keyphrase in the document and makes the extraction process easier and faster.

There are different ways to aggregate data such as-using semantic web approaches (linked data) and using natural language processing (NLP) approach.Machine learning approach is one of the approaches that NLP uses most commonly and it is also an effective approach for text analysis. Semantic web technology is very good to process web data. It has component like- RDF(Resource Description Framework), Web Ontology Language (OWL) and Extensible Markup Language (XML) The most famous content management systems are using semantic web approach.

On the other hand, natural language processing and machine learning approaches extract keyword from a learned model and training document creates the model for identifying keyphrases. One of the machine learning approach is Naive bayes. MAUI[18] [19], KEA [18] and GenEx [20] etc keywords extraction tools based on Naive bayes algorithm. The best thing of this algorithm is very efficient and fast as it uses pre-computed terms from a text. We can also use approaches like word co-occurrence statistical information, algorithm like TF*IDF for the keyword extraction. In addition to this, two other types of keyword extraction process PIC can follow such as free indexing and indexing using a controlled vocabulary. In free indexing, for keyword extraction does not need any vocabulary. Whereas,controlled vocabulary based indexing extracts the keywords or terms from defined vocabulary or thesauri.

4.5. Filtering data

Another issue is how to filter information once the content analysis or keyword extraction is done? Why filtering is important? Filtering is important to find similar content in PIC and to create one single view according to the users filtered input or selected terms or keyphrases. We need to figure out for our personal information center how to link different pieces of information or content using different approaches to find the similarity among them. Different filtering approaches like content based filtering or collaborative filtering can filter the similar content by

ranking or rating.

Finally as an additional step performance evaluation of the keyword extraction can be done by calculating different measure such as precisions, recall and F-measure.

So, we need to find a way to collect information from web and to process those information to extract keywords from the collected information. We need to find out the approaches, algorithms and tools to process raw information from the web and to extract important keyword from a document without users interaction. Finally, find similarity of different content so that users can see their personalized web content by filtering or selecting one term or keyphrases. Therefore, it is also important to define a tools to find out similar content in PIC.

5 Decisions

After analyzing the problems in the previous section it is clear that in PIC a user defines the sources of the data and what types of the data content (e.g.-email, calendar event or news items) they prefer. Then PIC completes all the processing automatically. If we go more deep inside the technical problems we find out the basic requirements of the problem. Having analyzed the task, we have discovered the basic requirements of PIC. These requirements are:

- Collecting information using a system or tools .
- Processing information and saving the results in the local machine in a specific format for example text file.
- Extracting the important key phrases from a text file and saving the extracted keyword in a new file using a tools or algorithm.
- Creating a vocabulary with extracted keywords and finding a way to match the terms of the vocabulary to tag or filter the similar terms content to create personalized web content view.

6 System Design

To meet the discussed requirements and to fulfill the users need for PIC, we designed a system which is mainly consists of an importer, an analyzer and a filtering system. To understand the overall system design we showed them in a layered approach. The first layer is content creation layer and an importer imports the data from multiple forces and create content for PIC. The second layer is extraction layer. the third layer is filtering and finally presentation layer.

An analyzer analyzes the content using different keyphrase extraction algorithms. The third layer implements different filtering algorithms such as collaborative or content based or any self learning algorithms to create a personalized system. Finally, the presentation layer presents different kinds of presentation for different device such as mobile, picture frame and PDA. For importing information PIC will be able to add, update or delete information from different sources at any time and users defined the information sources according to their choice but type of data will be from rss feeds, weather information, email message, calendar events, even camera image or information from local drive etc.

The core activity of the overall system design according to design artifacts is shown in Figure-6.1 and the overall system design shown in Figure-6.2.

6.1. Content creation layer

For information extraction layer at first PIC collects the information from the users defined source for example- internet and local storage. Then processes the information according to our discussed requirement in the previous chapter. For this layer an importer is develop to fetch, parse and process information. If we go in details of importer then we can say that, an importer is a class with three method `fetch()`, `parse()` and `process()`. This class holds a pointer to the three plugins `Fetcher`, `Parser()`, and `Processor()`. These are abstract classes and they define functionality or methods for them and these methods are called during the implementation. Once the method `processor()` is implemented a new node is created. But before creating a node, a processor maps the source of the information into a target information source from an initial source. Once the node is created it will be shown on the UI according to target information. the classes are shown below Figure-6.3

The whole importer system should look like below: figure-6.4

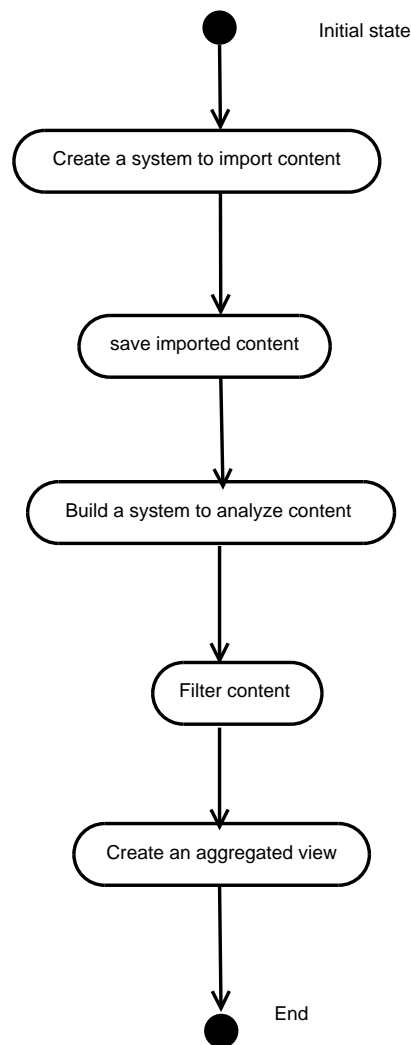


Figure 6.1.: Overall activity of system design

6.2. Extraction layer

Once the node is created an analyser does the real magic of keyword extraction. There are many algorithms to extract keywords. But analyzer follows the procedure indicated in collaboration diagrams. Figure-6.5 indicates how it extracts keyphrases in the system.

The collaboration diagram states that an analyzer should have a class named TopicExtractor() that extract the keyphrases from a learned model. So, there will be a ModelBuilder() class that build a model from node in the first layer. As analyzer consists of two classes therefore the class diagram for extraction layer is figure-6.6

From the figure-6.1 activity diagram we see that as TopicExtractor class loads a model first from a learned model to extract keyphrases. The detailed information of TopicExtractor and ModelBuiler class is mentioned in the figure-6.6 class diagram. ModelBuilder class builds a model saves the model for TopicExtractor.

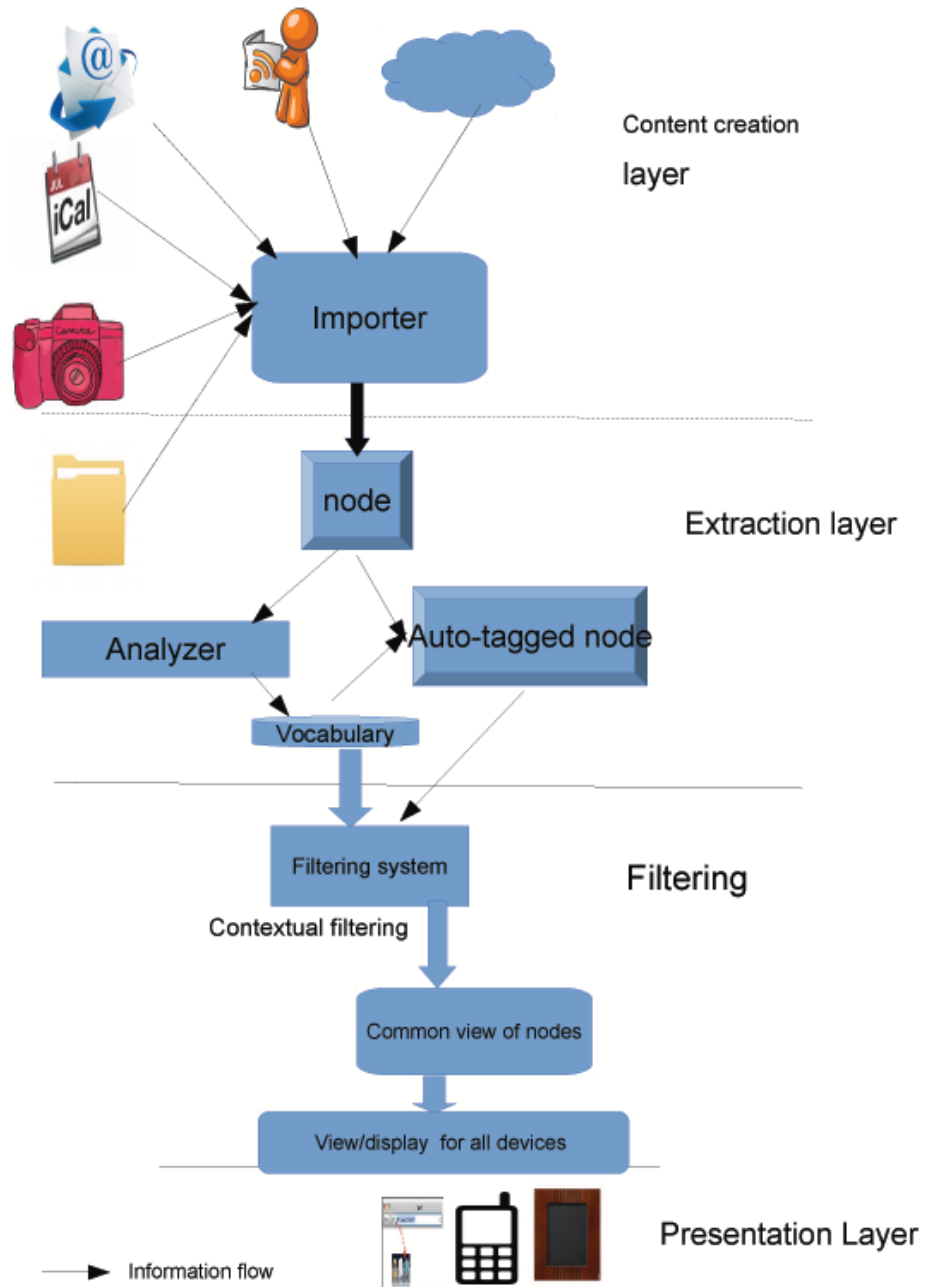


Figure 6.2.: System design

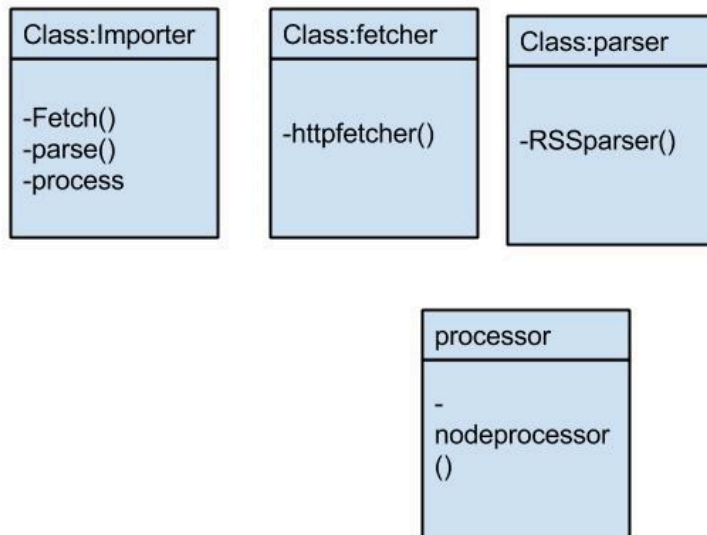


Figure 6.3.: Importer classes

So, after extracting keyphrases our system builds a vocabulary from keyphrases extracted by TopicExtractor class. PIC has an internal module with functionality to create automatic vocabulary terms from the extracted keyword. This module is designed in a way so that it can fulfill the requirements of auto tagging content. But to implement the module PIC creates some rules to autotag term from a text. So, the behavior of the system for this extraction layer is designed as follow using a state diagram Figure-6.7

6.3. Filtering Layer

Filtering layer looks for filtering algorithms to design a filtering system for PIC. To design an efficient filtering system, PIC follows content based filtering strategies to match the content with other content by looking at the similar text pattern or keywords. PIC matches the vocabulary terms with the already tagged content to find the similarity of the content and then show the content automatically in a common view. In other words, If any overlapping occurs then this filtering system of PIC creates a common view of nodes with similar or overlapped terms. Some of the content management system is using content recommendation module as their filtering strategy.

6.4. Display Layer

This layer presents the personalized content in the filtering layer in a way so that PIC can be accessible from multi-device platform. In this layer PIC presents different types of views (display) for different device such as slideshare view or image view for picture-

The screenshot displays the 'Importer' settings interface. It is organized into four main sections, each with a title and a 'Change' link:

- Basic settings**: Includes 'Attached to: Feed' with a 'Settings' link, 'Periodic import: every 30 min', and 'Do not import on submission'.
- Fetcher**: Includes 'HTTP Fetcher' with a 'Settings' link and the description 'Download content from a URL.'.
- Parser**: Includes 'Common syndication parser' with the description 'Parse RSS and Atom feeds.'.
- Processor**: Includes 'Node processor' with 'Settings' and 'Mapping' links, and the description 'Create and update nodes.'.

Figure 6.4.: Importer

frame¹.

The system designed for the PIC is meant to be scalable, user friendly and data privacy concerned. User has the ability to modify and organized the content created in the system. So, it is easily adaptable by the user. Though in the beginning our system is looking for a filtering algorithm which is same as the content based filtering but later according to the user requirement it is also possible to use another filtering approach. The PIC system is designed for users to personalize their web content management. All the processing steps within the system are done automatically with data from the users defined information sources without requiring user action. The system first categorizes these sources based on the users preference and then created a common view. One of the main advantages of this system is it grants the user the ability to add data from their camera or mobile devices.

¹ <http://www.usa.philips.com/c-m-so/digital-photo-and-video>

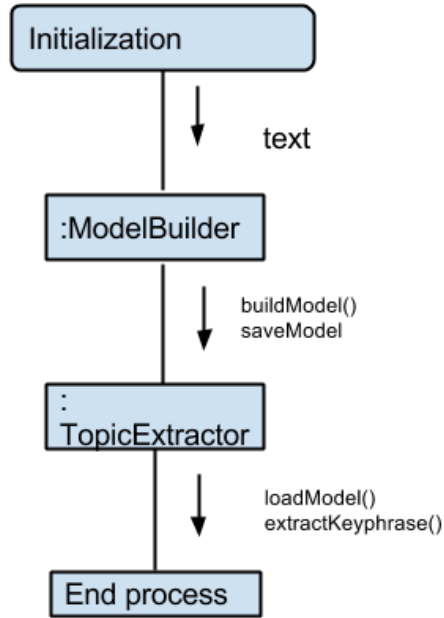


Figure 6.5.: Collaboration diagram of the extraction process

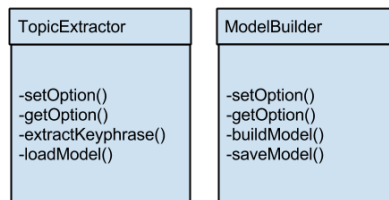
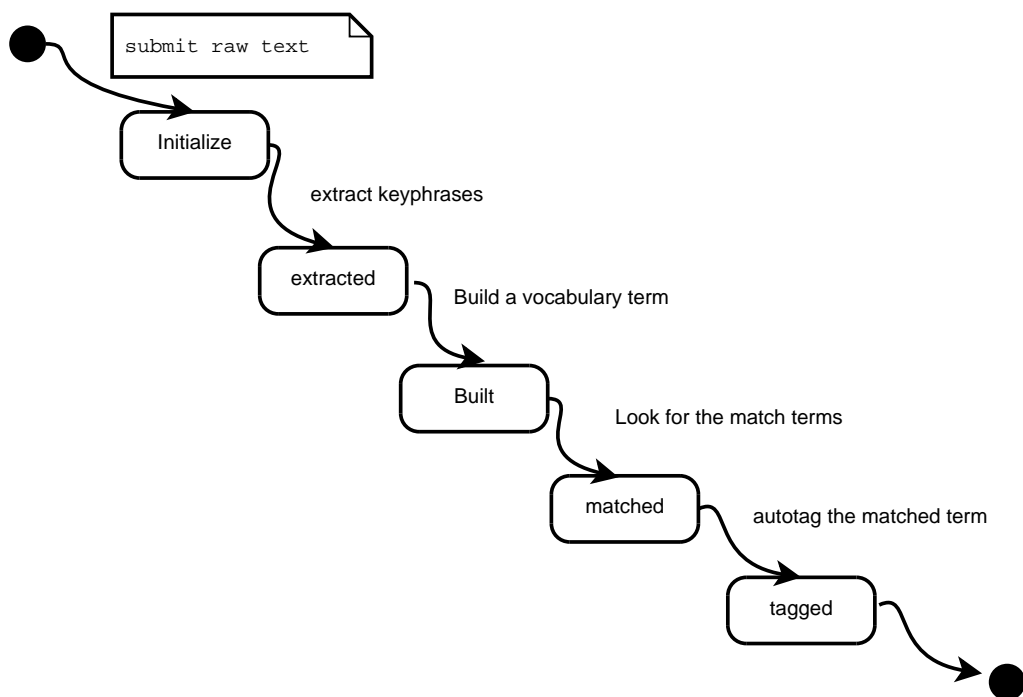


Figure 6.6.: Class diagram of the extraction process

**Figure 6.7.:** State Diagram

7 Deployment

The prototypical system of PIC is deployed on a linux machine (Laptop) with a processor of intel core i5 and also on raspberry pi with a linux kernel based OS called raspbian. At first LAMP server(Linux-Apache-MySQL-PHP) server is configured on Ubuntu 13.10. For LAMP configuration Apache 2, PHP 5 and MySQL 5.0 is configured. Then we deployed drupal version 7 content management system and MAUI 1.2 keyword extraction software on top of the LAMP server to implement the designed system. The software included in Maui are KEA, Weka, Jena and Wikipedia Miner. But for this thesis implementation phase we are only using KEA and Weka with Maui.As LAMP server has a local php-mysql database therefore drupal uses this as a central database to store, retrieve and update content. All drupal modules are also stores in the php-mysql database. The deployment architecture is shown in figure-7.1

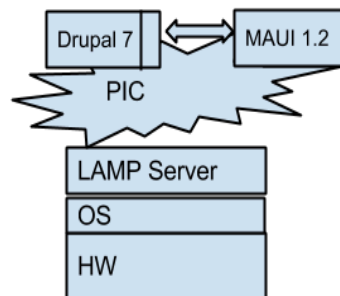


Figure 7.1.: Deployment Diagram

8 Implementation

We implemented the overall system of PIC using drupal 7 content management system with its extendable contributed modules and Maui 1.2, a JAVA based keyword extraction tool. From the Figure-8.1 we see that the contributed module plays an important role to implement our designed system.

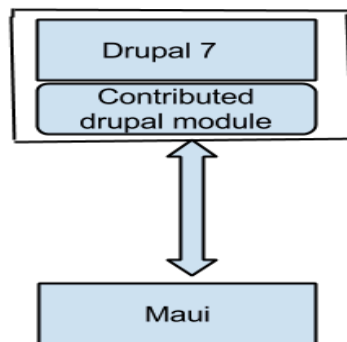


Figure 8.1.: Maui and Drupal 7 communication using drupal custom module

In this thesis, a contributed drupal 7 module is used to connect the functionality of MAUI system with drupal system. The main benefits of using modular drupal content management system is it is modifiable easily and its extendability of the modules also increases the functionality of the system. Therefore, to meet the designed system requirements we have decided to create contributed drupal modules that run the MAUI keyword extraction tools from drupal system and grab the extracted keyphrases from MAUI local file system and transfer them to the drupal system. The activity diagram figure-8.2 indicates the overall activity implemented in order to build the entire system..In the following sections of this chapter we will talk about the whole implementation in details.

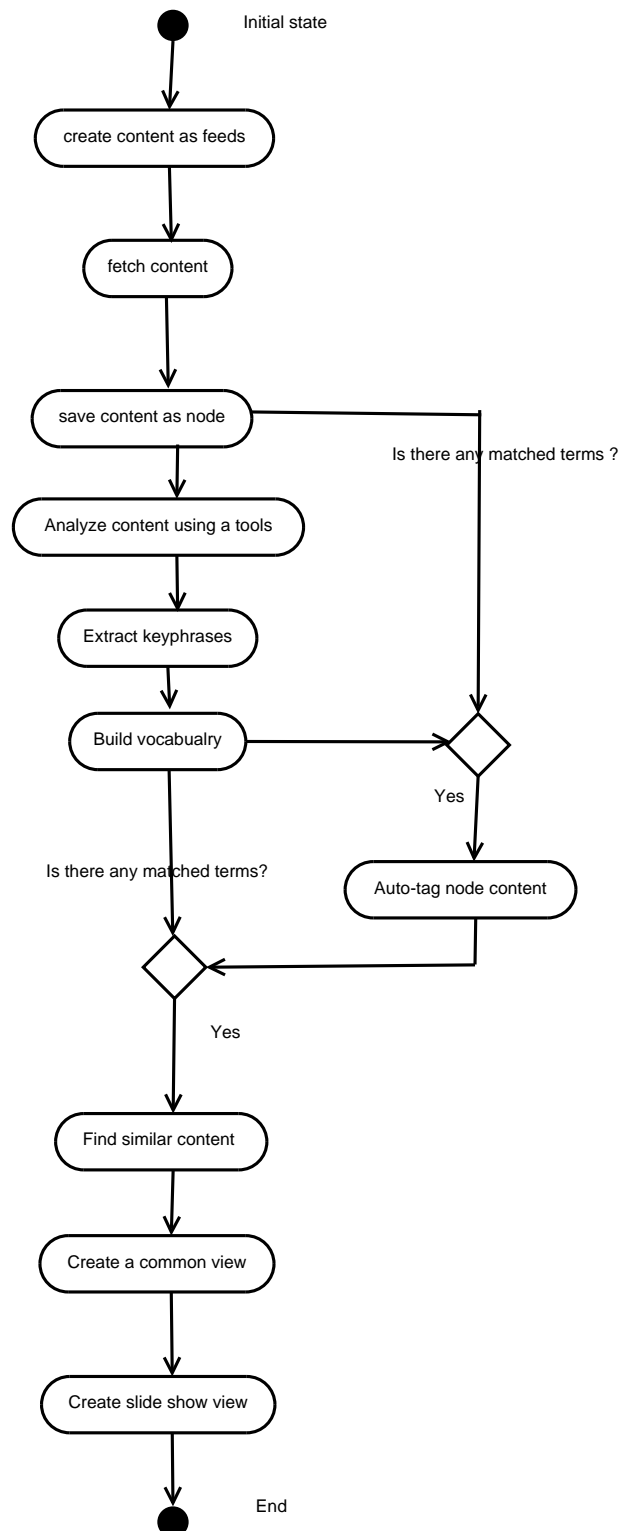


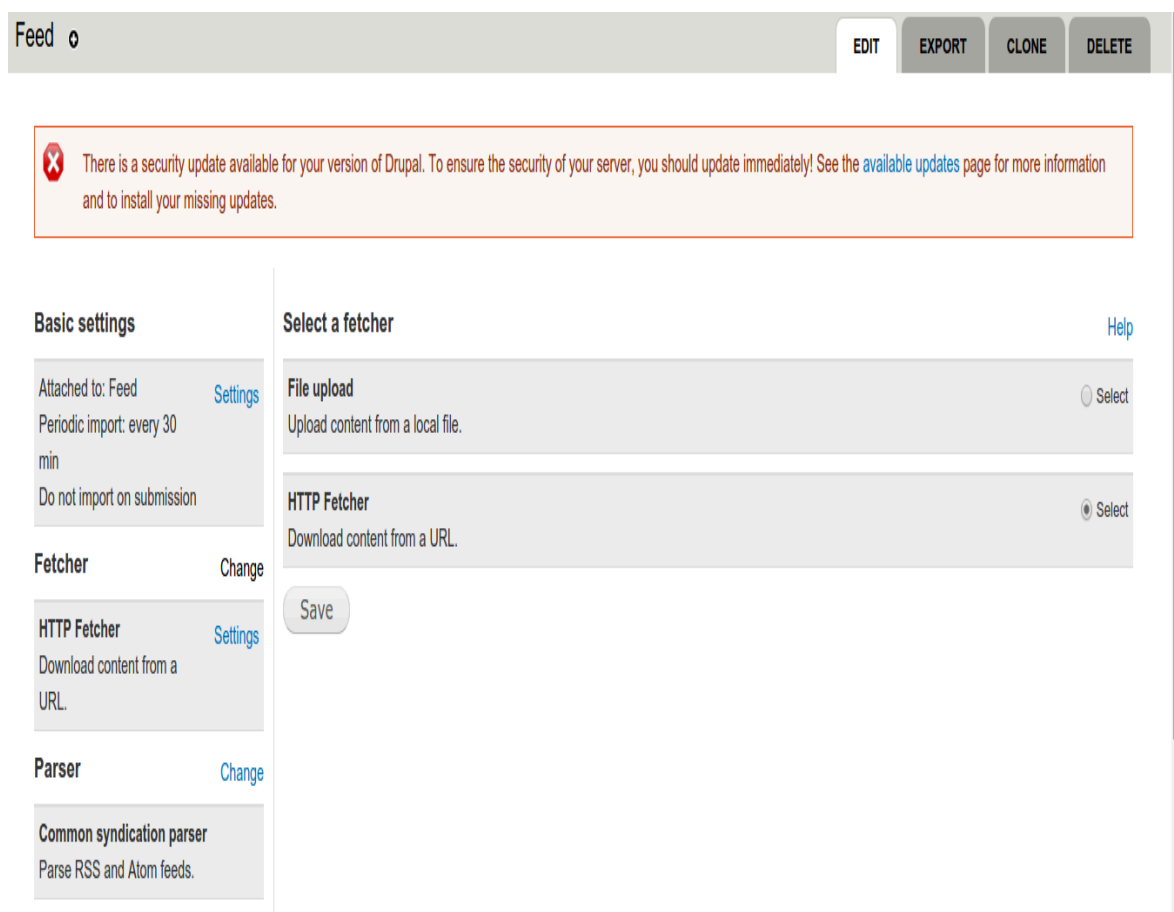
Figure 8.2.: Activity diagram for PIC system implementation

8.1. Content creation

We create content from different sources as discussed earlier and sources are defined by the users. For our implementation we create all content as feeds. Feed is the body of the data or content. Feeds has title and feeds items.

8.2. Collecting information

Once the content is created an importer imports the content using three different plugins: Fetcher, Parser, processor. Fetcher downloads or stores or retrieves the the feed from a specific location or path or from a url so that parser can parse the feed to a feed items. Feeds items are auto generated once a feeds is created. Figure-8.3 and figure-8.4 shows fetcher and parser in PIC. From the Figure-8.3 we can see that fetcher fetches two types of information. First option is, it uploads content from a local file. Second, it uploads content from url using HTTP Fetcher. PIC uploads content using an HTTP Fetcher since we are collecting information as a feed.



The screenshot shows the 'Feed' configuration page in Drupal. At the top, there are buttons for 'EDIT', 'EXPORT', 'CLONE', and 'DELETE'. A security update notification is displayed, stating: 'There is a security update available for your version of Drupal. To ensure the security of your server, you should update immediately! See the available updates page for more information and to install your missing updates.'

The main configuration area is divided into two columns. The left column contains settings for the feed's basic settings, the fetcher, and the parser. The right column is titled 'Select a fetcher' and contains two options: 'File upload' and 'HTTP Fetcher'. The 'HTTP Fetcher' option is selected.

Basic settings

- Attached to: Feed [Settings](#)
- Periodic import: every 30 min
- Do not import on submission

Fetcher [Change](#)

- HTTP Fetcher** [Settings](#)
Download content from a URL.

Parser [Change](#)

- Common syndication parser**
Parse RSS and Atom feeds.

Select a fetcher [Help](#)

- File upload** Select
Upload content from a local file.
- HTTP Fetcher** Select
Download content from a URL.

[Save](#)

Figure 8.3.: Fetching information in PIC

From figure-8.4 we can see that there are many options of parsing such as common spec-

Select a parser Help

Common syndication parser
Parse XML feeds in RSS 1, RSS 2 and Atom format. Select

CSV parser
Parse data in Comma Separated Value format. Select

OPML parser
Parse OPML files. Select

Sitemap parser
Parse Sitemap XML format feeds. Select

Figure 8.4.: Parsing information in PIC

ification parser, CSV parser, OPML parser and sitemap parser. But in PIC the parser parses informations by using common specification parser. This common specification parser can parse XML feeds in RSS and atom format. So, after selecting the common specification parser we save the parser option. The parser parses the feeds to process it through the processor. Processor maps the source feed item to a targeted feed item. Figure-8.5 shows an example of mapping procedure PIC uses.

From the figure-8.5 we can see that the mapping procedure can source feed item which will be mapped into a target feed item. From our figure the mapping is as follows:

- Title maps to Title
- Published date (timestamp) maps to Published date(created)
- Item GUID(guid) maps to GUID(guid), which is an unique field
- Item url to URL and so on.

Therefore, once the mapping is done according to the mentioned discussion the processor saves the mapped feed item as a node. More precisely, the processor creates or updates a nodes. Once node is created it is stored in the database with a unique node id or **nid**.

8.3. Keyphrase extraction

As we are using an external tools, Maui to extract keyphrases from the content, therefore a custom drupal module (MAUIKeywordExtractor.module) sends the nodes content to Maui system. Our custom module sends the content to a training document set and saves it in a specific location to process the training document set. Basically, drupal sendd the content to maui as a text file. This text file is a input document for the keyphrase extrac-

tion. This MAUIKeywordExtractor.module has a great contribution for keyphrase extraction but it cannot work alone. It works as an api of content analysis main module. Now, there are lots of processing happens inside Maui system. Maui uses KEA, keyphrase extraction algorithm to extract keyword. Our designed system extract keyphrases automatically using KEA algorithm from the input training document set. The followings are the steps to process the input file for keyphrase extraction.

1. First, Maui splits the input text into tokens and then selects candidate phrases from the input text file. Candidate phrases are the most important keyphrases in the document. Therefore candidate selection is very important step for keyword extraction. KEA reads each line of the input file to identify candidate phrases. To extract the keyphrases at first KEA extract n-grams from the input file, removes all stop-word and stemming. Maui follows a stop words list containing 425 words and these word list are syntactically an adverbs, adjectives, anomalous verbs, conjunction, articles, prepositions, pronouns etc. For stemming Maui uses porter stemming algorithm. Maui uses some rules to select candidate phrases. These rules are-
 - stopwords are removed from the beginning and ending of a collected n-grams
 - and n-grams should have a maximum word length of five words.

These rules make the training and extraction process faster. As we are using free indexing therefore, there is no matched thesaurus terms in the collected the n-grams.

Final step of candidate phrases selection is to calculate some features for each candidate. These features are TF*IDF, first occurrence, length of the keyphrases and keyphraseness. According to Maui algorithm our PIC computes following features for the candidate phrases selection:

- TFxIDF : Set this feature to True
 - First occurrence : Set it to a minimum limit of 1
 - Length : Minimum phrase length is set to 1 and maximum is set to 5.
 - Keyphraseness: Minimum value is set to 0.01
2. Once the candidate phrases are identified maui builds a model using Naive Bayes machine learning approach from the input document. For model building Maui uses Weka¹ machine learning toolkit. The input document is known as training document set according to Maui algorithm. Finally, Maui extracts keyphrases in a new file using the model built from the training document set and saves the extracted keyphrases in a new file with a extension of .key.
 3. After keyphrases extraction our module MAUIKeywordExtractor.module calls the extracted keyword into the drupal system and creates taxonomy vocabulary terms

¹<http://www.cs.waikato.ac.nz/ml/weka/>

Mapping for Node processor [Help](#)

Define which elements of a single item of a feed (= Sources) map to which content pieces in Drupal (= Targets). Make sure that at least one definition has a *Unique target*. A unique target means that a value for a target can only occur once. E. g. only one item with the URL <http://example.com/content/1> can exist.

[Show row weights](#)

SOURCE	TARGET	TARGET CONFIGURATION	
+ Title (title)	Title (title)	Not used as unique.	<input type="checkbox"/> Remove
+ Published date (timestamp)	Published date (created)		<input type="checkbox"/> Remove
+ Item GUID (guid)	GUID (guid)	Used as unique .	<input type="checkbox"/> Remove
+ Item URL (link) (uri)	URL (uri)	Not used as unique.	<input type="checkbox"/> Remove
		Text format: <i>Full HTML</i>	

Figure 8.5.: Mapping for node processor

The screenshot shows a Drupal 7 administration page with a 'Content Analysis Results' pop-up window. The window is titled 'Content Analysis Results' and has a 'Close Window' button. It displays the 'MAUI Keyword Extractor' results for a report. The results are listed in a table with a 'Report' column and a 'Keyword' column. The keywords are:

- Iberian Peninsula in southwestern mainland is bordered
- Western Europe and the European member of the United
- Economic Co
- lez de la Gomera
- Spain is a democracy organized
- United Nations
- 15th century
- unified

Figure 8.6.: Pop up view of the Keyphrases extraction using Maui

from the extracted keywords. For this research study we selected drupals default vocabulary named **Tags**. Figure-8.6 shows the extracted keywords in an pop up window. From the figure we can that the maximum length of keyphrases is not more than five word length. So, once these keyphrases are extracted an automatic vocabulary will be created with the extracted keyphrases or terms. And the custom module MAUIKeywordExtractor.module has the main contribution to create this automatic vocabulary terms.

The algorithm of communication between custom module and Maui is mentioned below:

- a) **Step 1:** Custom module sends the node content to Maui system and save it as a text format.
- b) **Step 2:** Custom module makes a function call to execute the keyphrase extraction command from its own drupal environments.
- c) **Step 3:** Maui processes the document and creates a training document set using algorithms inside Maui.Maui builds model and extract keyphrases by learning the model and save the extracted keyphrases in to a new file with extension .key
- d) **Step 3:** Again custom module opens the extracted keyphrase file and reads each and every line.
- e) **Step 4:** Finally, create an automatic vocabulary from the keyphrases extracted by Maui.

The module code is mentioned in 8.1.

Listing 8.1: Custom Maui keyword extraction module

```

1 <?php
2
3 // $Id: mauikkeywordextractor.module,v 1.6 2010/05/09 19:55:26
   tomdude48 Exp $
4
5 /**
6  * @file
7  * An example content analyzer using the Content Analysis API
8  */
9 function mauikkeywordextractor_menu() {
10   $items = array();
11
12
13
14   return $items;
15 }
16
17 /**
18  * Implementation of hook_contentanalysis_analyzers()
19  * register contentanalysisexample with contentanalysis analyzers
   registry
20  */
21 function mauikkeywordextractor_contentanalysis_analyzers() {
22   $analyzers['mauikkeywordextractor'] = array(

```

```

23     'title' => t('MAUI Keyword Extractor'),
24     'module' => 'mauikeywordextractor',
25     'callback' => 'mauikeywordextractor_analyzer',
26     //'form elements callback' => '
        mauikeywordextractor_analyzer_form_elements',
27     //'node form submit callback' => '
        mauikeywordextractor_node_form_submit',
28     'weight' => -5,
29 );
30 return $analyzers;
31 }
32
33
34
35
36 /**
37  * Implementation of hook_analyzer() via custom define callback
38  *
39  * Performs the analysis.
40  * callback is defined in hook_contentanalysis_analyzers ['callback
    ' ]
41  *
42  * @param unknown_type $context
43  *   Array context format defined by contentanalysis.module
44  * @param unknown_type $analysis
45  *   Array analysis format defined by contentanalysis.module
46  * @param unknown_type $params
47  *   Array customer defined paramters
48  */
49 function mauikeywordextractor_analyzer($context, $analysis, $params
    ) {
50
51
52 // call maui implementation with the parameters $context['body']
53
54
55
56 $terms = mauikeywordextractor_autokeyword($context, $analysis);
57
58
59 $rows = array();
60 $header1 = array(
61     array('data' => t('Term')),
62     //array('data' => t('Relevance')),
63 );
64 if (is_array($terms)) {
65     foreach ($terms as $v) {
66         $rows[] = array(
67             "<span class=\"kwresearch_keyword\">" . $v . "</span>"
68             // "<span class=\"rules_autotag\">" . $v . "</span>"
69         );
70     }

```



```
71 }
72 if (!$rows) {
73     $rows[] = array(array(
74         'data' => t('No keywords available.'),
75         'colspan' => count($header),
76     ));
77 }
78
79
80
81 $out = theme('table', array('header' => $header, 'rows' => $rows)
82     );
83 $analysis['content'][] = contentanalysis_format_content($out, -1)
84     ;
85 // send out content into a text file
86
87 if (isset($_POST['analyzers'])) {
88     $content = $_POST['body'];
89     $file = fopen("/var/www/dru7/sites/all/modules/contentanalysis/
90         Maui1.2/data/automatic_tagging/train/merge/content.txt", "w
91         ");
92     fwrite($file, $content);
93     fclose($file);
94     // print_r(error_get_last());
95 }
96
97 return $analysis;
98 }
99
100 /**
101  * @todo Please document this function.
102  * @see http://drupal.org/node/1354
103  */
104 function mauikkeywordextractor_autokeyword($context, &$analysis =
105     NULL) {
106     /*
107     * Run maui from drupal
108     */
109     exec("cd /var/www/dru7/sites/all/modules/contentanalysis/Maui1.2/
110         && touch /tmp/testfile.txt && java -cp /var/www/dru7/sites/
111         all/modules/contentanalysis/Maui1.2/bin:/var/www/dru7/sites/
112         all/modules/contentanalysis/Maui1.2/lib/* maui.main.
113         MauiTopicExtractor -l /var/www/dru7/sites/all/modules/
114         contentanalysis/Maui1.2/data/automatic_tagging/train/merge/ -
115         m test -v none");
116     unlink($filename);
117
118     // get te file content into drupal
```

```

112 $filename = "/var/www/dru7/sites/all/modules/contentanalysis/
      Mauil.2/data/automatic_tagging/train/merge/content.key";
113 file_get_contents($filename);
114 $terms = array();
115 $handle = fopen($filename, "r"); // handle input files ,reading
      input file
116 $vid = taxonomy_vocabulary_machine_name_load('tags')->vid;
117 while ($line = fgets($handle)) {
118   taxonomy_term_save((object) array(
119     'name' => $line,
120     'vid' => $vid,
121   ));
122   array_push($terms, $line);
123   }
124
125
126
127   fclose($handle);
128   unlink($filename);
129
130
131   return $terms;
132 }

```

4. Now, we use another module rules autotag to tag the content automatically with MAUIKeywordExtractor.module. To implement the rules autotag module we create a rules to autotag the text or content. Then, autotag rules module matches the vocabulated terms with the text and tag the matching terms from the text file automatically. The listing 8.2 shows the created rules for autotag content and the source code 8.3 of the rulesautotag.module.

Listing 8.2: Rules for autotaging content

```

1
2 { "rules_autotagging" : {
3   "LABEL" : "autotagging",
4   "PLUGIN" : "reaction rule",
5   "OWNER" : "rules",
6   "REQUIRES" : [ "rules", "rules_autotag", "feeds" ],
7   "ON" : { "node_presave" : [], "feeds_import_feed" : [] },
8   "IF" : [
9     { "entity_has_field" : { "entity" : [ "node" ], "field" : "
      field_tags" } }
10  ],
11  "DO" : [
12    { "rules_autotag_extract" : {
13      "USING" : { "text" : [ "node:body:value" ], "vocabulary" :
      "tags" },
14      "PROVIDE" : { "extraced_terms" : { "extraced_terms" : "
      Extracted terms" } }
15    }
16  ],
17  { "data_set" : { "data" : [ "node:field-tags" ], "value" : [ "

```

```

    extraced-terms" ] } }
18 ]
19 }
20 }

```

Listing 8.3: Custom Rules Autotag module

```

1 <?php
2
3 /**
4  * Extract terms.
5  *
6  * @param $text
7  *   The text being parsed by the extractor.
8  * @param $vocabulary
9  *   Vocabulary defining terms which should be extracted.
10 *
11 * @return
12 *   An array of tids.
13 */
14 function rules_autotag_extract($text, $vocabulary) {
15
16 $file = fopen("/var/www/dru7/sites/all/modules/contentanalysis/
17   Maui1.2/data/automatic_tagging/train/merge/content.key", "w");
18   fwrite($file, $text);
19   fclose($file);
20
21 exec("cd /var/www/dru7/sites/all/modules/contentanalysis/Maui1.2/
22   && java -cp /var/www/dru7/sites/all/modules/contentanalysis/
23   Maui1.2/bin:/var/www/dru7/sites/all/modules/contentanalysis/
24   Maui1.2/lib/* maui.main.MauiTopicExtractor -l /var/www/dru7/
25   sites/all/modules/contentanalysis/Maui1.2/data/
26   automatic_tagging/train/merge/ -m test -v none");
27
28 $filename = "/var/www/dru7/sites/all/modules/contentanalysis/Maui1
29   .2/data/automatic_tagging/train/merge/content.key";
30 unlink($filename);
31
32 // get te file content into drupal
33 $filename = "/var/www/dru7/sites/all/modules/contentanalysis/
34   Maui1.2/data/automatic_tagging/train/merge/content.key";
35 //file_get_contents($filename);
36 $terms = array();
37 $handle = fopen($filename, "r"); // handle input files ,reading
38   input file
39 $vid = taxonomy_vocabulary_machine_name_load('tags')->vid;
40 //$vid = taxonomy_vocabulary_machine_name_load($machine_name)->
41   vid;
42
43 while ($line = fgets($handle)) {
44   taxonomy_term_save((object) array(
45     'name' => $line,
46     'vid' => $vid,

```

```

37 ));
38 }
39 fclose($handle);
40 unlink($filename);
41
42
43
44 $extracted_tids = array();
45 $terms = rules_autotag_get_term_names($vocabulary);
46
47 $text = rules_autotag_clean_text($text);
48 $text_tokens = array_flip(rules_autotag_split_text($text));
49
50 $matchings_term_splits = array_intersect_key($terms, $text_tokens
51 );
52 // Loops over all matched splits and checks if a term name
53 // consists
54 // of multiple splits. If so, an additional text parsing for the
55 // whole
56 // term name is performed.
57 foreach ($matchings_term_splits as $results) {
58   foreach ($results as $result) {
59     $tid = $result['tid'];
60     if (!in_array($tid, $extracted_tids)) {
61       if ($result['splitted']) {
62         if (strpos($text, $result['original_term_name']) !==
63             FALSE) {
64           $extracted_tids[] = $tid;
65         }
66       }
67     }
68   }
69 }
70 return $extracted_tids;
71 }
72
73 /**
74  * Returns an array of terms, keyed by splitted term names.
75  *
76  * The structure can be modified with
77  *   hook_rules_autotag_terms_alter()
78  * implementations.
79  */

```

8.4. Filtering

To find out the similar content or to filter out the user preferred content and to see it in a single view or display format another drupal 7 module named similarity by terms is

installed. Here, view means display or a web page, user interface or block. This module filters out all the similar contents when a user visits a node where node contents are already tagged. This module is following content based filtering algorithm and in drupal it is known as content recommender module. Though there are more drupal content recommender modules such as relevant terms, similar objects and apache solr. But similar by terms meet our research requirements to create a common view with the similar terms. After importing the content an importer saves it as a node. Therefore, we filter out nodes according to their node id using contextual filtering. Contextual filter also known as a view(page) arguments.

8.5. Presentation

Finally, for presentation purpose we create a slideshow view. In this view, we can see all our aggregated content in a slide show. We can view our customized content from a device like picture frame using slide show view or image. To implement slideshow view drupal slideshow view module is needed. Basically, it is a view or presentation format that presents the aggregated content all together.

8.6. Protocols used for this implementation

The protocols used for the implementation of PIC is listed in the table-8.1.

Type of the data	protocols/libraries
Email	IMAP, POP3
Calendar	ical

Table 8.1.: Protocols used in PIC

8.7. Modules and features

We install a list of drupal 7 modules to implement PIC which is mentioned in table-8.2.

Modules	Purpose
Views	To display, the page
View Slideshow	To create a slideshow views or display
Feeds	Create content as a feeds and importing feeds content
Content analysis	To analyze the content in different ways
Keyword analysis	To analyze extracted keyphrases
Rules	create rules for, multiple purpose
Rules Autotag	To tag content automatically
Similar by terms	create a common view for similar content

Table 8.2.: Installed drupal 7 modules for PIC

8.8. Final display

Figure-8.7 shows a common display for all similar nodes. From the figure we can see that all the nodes that have autotag keywords for example -"centre" are shown in a block view

8.9. Challenges for Implementation

The main challenge for implementation was to select keyword extraction tools that will run on local machine due to data privacy issues. As we are using users sensitive data the privacy of the data was of great concern and instead of using drupals own content analysis API we decided to use an external software, Maui to implement this thesis. But as Maui is a java based software and drupal is a content management framework based on php, connecting Maui and drupal was a another challenge in the beginning of this master thesis implementation. Finally, we solved this issue by developing a contributed module called MAUIKeywordExtractor.module.

[Home](#)

Navigation

- ▶ [Add content](#)
- ▶ [Import](#)

Related Terms

- [Workers suspended over 'data breach'](#)
- [VIDEO: Ebola charity: 'No stone unturned'](#)
- [VIDEO: Inside an aeroplane 'black box' lab](#)
- [Italy](#)
- [Inmates 'drunk on hooch' in disorder](#)
- [Workers suspended over 'data breach'](#)
- [VIDEO: Ebola charity: 'No stone unturned'](#)

VIDEO: Inside an aeroplane 'black box' lab

[View](#) [Edit](#) [Log](#)

Submitted by Anonymous (not verified) on Mon, 01/12/2015 - 08:07

Richard Westcott reports from an investigation centre where flight recorders, like the one being examined in Indonesia, are listened to and inspected.

Tags:
[centre](#) [centre](#)

Example Description:
Richard Westcott reports from an investigation centre where flight recorders, like the one being examined in Indonesia, are listened to and inspected.

[Add new comment](#)

Figure 8.7.: Similar node in a common view

9 Evaluation

In this chapter we are going to evaluate the research questions of this thesis. Then we also evaluate drupal 7 for the developed system. For the implementation we used raspberry pi and desktop. Therefore, at first we are going to evaluate the hardware performance of the prototypical system.

9.1. Performance of the hardware

The prototypical system is running successfully on lenovo x86.64 machine architecture. We also tried to run it on raspberry pi and it worked well but the performance was bit slower than lenovo intel 5i.

9.2. Data collection using our designed system

To meet the thesis requirement of data collection we build a feed importers that can fetch information from the multiple sources. The importer processes data as a node according to our system design. However, our built system processes data as a node but also processes data in different formats such as taxonomy term processor, user processor, or workflow processor to create and update workflow. Our designed importer does not only fetch data from http url but also from a local storage or directory. Therefore, we can say that the designed feeds importer for fetching information solved the first research challenge properly and takes our system design one step forward to fulfill the next requirements of the system design.

9.3. PIC a system that aggregate data

In PIC data is aggregated in many ways. At first after importing data the importer saves it as a node content. Then it extracts keyphrases using MAUI keyword extraction algorithm. Maui has different performance measures such as precision, recall and F-measure to evaluate the performance of the keyword extraction. But before talking about the measure if we look at the keyphrase extraction then we see Maui has a collections of algorithm

No. of training document set	Precision	Recall	F-measure
15 training Set	88.67 +/- 10.6	89.83 +/- 9.28	89.83 +/- 9.28
Single training set	100 +/- 0	100 +/- 0	100

Table 9.1.: Performance evaluation of keyword extraction

The upper value is the average precision and recall. The lower value is standard deviation value

to make the keyphrase extraction perfect. For example- extracting n-grams for candidate selections and for that porter stemming algorithm is used and even all stops are also removed in order to extract n-grams. Finally, it features calculation such as TF*IDF, first occurrence and keyphraseness etc makes the extraction process more robust to select most significant keyphrases only.

Finally, the calculation of precision, recall and F-measure evaluates the performance of keyword extraction process mathematically. At first the evaluation was performed using 15 documents manually created by the user with different email messages, journals and newspaper articles. The F-measure of the keyword extraction was 89.25%. We also evaluated the F-measure for single document created by PIC analyzer and it was 100%. Table-9.1 indicates the value of F-measure, precision and recall after keyphrases extraction.

From the table we can see that for a single training document the precision and recall value is 100% in compared to 15 training document sets.

9.4. Filtering system for personalized web content

To evaluate the filtering system we went through a node which was already tagged and it had tagged keyword events. Once we clicked the tagged keyword we found a block view with a list of events related to the tagged keyword events. Figure-9.1 ...shows that our developed system filters out all kind of events such as music , video etc.

9.5. Presentation of PIC

Presentation of PIC is an important issue especially for the device like picture frame. To solve this problems PIC creates different types of views such as common view using filtering system, slideshow view using slideshow views modules. This slideshow view is very appropriate for device like picture frame.

9.6. Data privacy of PIC

If we evaluate the data privacy of the system then we can say our built system is very safe. All the data processing is done inside the local machine. Therefore, users sensitive data for example email and calendar data remains private and secure.

The screenshot shows a Drupal node view. On the left, there is a 'Navigation' sidebar with links for 'Add content' and 'Import'. Below it is a 'Related Terms' sidebar listing various topics like 'Music in 2015: A look-ahead Did King John actually 'sign' Magna Carta?' and 'Spain's ex-king faces paternity suit'. The main content area features the title 'Is real Martin Luther King forgotten?' with 'View', 'Edit', and 'Log' buttons. Below the title, it states 'Submitted by Anonymous (not verified) on Mon, 01/19/2015 - 10:18'. The main text reads: 'As people across the US take part in remembrance events dedicated to Martin Luther King Jr., many are concerned his image has become too sanitised.' There are tags for 'King', 'King', 'events', and 'events'. An 'Example Description:' section repeats the main text. At the bottom, there is a form to 'Add new comment' with fields for 'Your name' (filled with 'admin') and 'Subject'.

Figure 9.1.: Similar node in a common view

9.7. Drupal CMS for PIC

Our designed system is transparent so we can easily see that importer is responsible for fetching data from multiple sources and creating a new node from the imported data. The analyzer is doing all kinds of extraction and analysis according to the system requirement. Filtering technology selection is also an important part of PIC system design phase. Therefore, using drupal contextual filter is a easy and simple way to create a common view according to the node id.

Using modular drupal content management system makes it easier to solve the problem of adding an external keyphrase extraction tools like Maui to the designed system. Drupal content management system was a great selection for the designed system to increase the functionality of the system and to meet the users requirements. In the beginning of this research we tried to evaluate different content management system with drupal. According to the evaluation matrix drupal fulfills almost all of the requirements to design our system. This evaluation matrix is enclosed with this thesis.

Therefore, the proposed research study solve all of the research questions mentioned to build a successful personal information center(PIC). The prototypical system, PIC is able to provide seamless facilities to the users by aggregated data from multiple sources with

a better performance.

10 Discussion

We are going to compare PIC with the relevant literature mentioned in the literature review. This chapter discusses a clear comparison among the reviewed literatures and the proposed research. In addition, similarity of the research study is also discussed.

In PIA [3] system the authors presented an agent based technology for their system. They used TF*IDF to extract keyword and collaborative filtering strategy. They do not mention anything related to n-grams extraction for pre-processing training document set, instead they use TF*IDF algorithm to extract keyphrases and to evaluate the accuracy of efficient retrieval and they use porter stemming algorithm and removed stop-words to process documents. But in our implementation the whole extraction process is done using Maui. Maui itself an individual tool to implement efficient keyphrase extraction. Maui has two main stages for keyphrase extraction. First, training and second extraction. Training consists of model building from candidate phrases. So, generating candidate consists of several steps. In candidate selection phase is a pre-processing steps where n-grams are extracted to create candidate keyphrases. Candidate selection is the first steps to process the input document. Then the final extraction begins once the training document set is process and a model is built from the training document set. All these steps makes the extraction process of PIC more accurate. In addition, PIA does not report any performance measure for their extraction process. But as PIC extracts keyphrases using Maui which has different measures such as precision, recall and F-measure. For our PIC the performance measure was 100% and extraction was done from a single document. For filtering PIC uses content based filtering and in implementation PIC is able to create an automatic view with the filtered keyphrases. Therefore, though PIC follows PIA framework but PIC has more advance algorithms to extract keyphrases and even it can create a common view from the filtering.

Other frameworks such as Fab [6] use both of the filtering strategies for efficient information retrieval. P-Tango [15] uses collaborative filter. PEA [5] uses evolutionary algorithm to find similarity of email content. Letizia [7] uses heuristics to predict which document is relevant for the user. Amalthea [16] is also an agent based system like PIA but uses evolution programming for filtering. In contrast, PIC uses more logical and robust tools and algorithm to extract keyphrases and to find similar contents. They even do not have any performance measure for their system.

Some frameworks use neural network based algorithm to generate training set but in their evolution they found that it is expensive, although they show good performance in

classification of documents. But training set generation of PIC is not expensive rather it is fast and robust due to extraction algorithm and model building strategy.

Our system extracts keyphrases using machine learning approaches mentioned in KEA [18] but with some extra features. Medelyan (2006) [19] mentioned that domain specific indexing increases the chances of better precision and recall value and in their study they showed 28.3% recall and 26.1% precision but PIC uses free indexing in its extraction algorithm with good results.

Semantic web based approaches also shows good results in information retrieval. OWIR [8] and PIRATES [9] are the systems that uses such approach. PIRATES presents a framework that can retrieve information and autotag contents. It also works the same way PIC works. PIC also auto tags content from the extracted keyphrases by creating an automatic vocabulary from the extracted keyphrases. Whereas, PIRATES extracts the keyphrases and use ontology domain to auto tag content from the extracted keyphrases.

From the discussion it is clear that PIC tries to integrate and adapt all of the above mentioned idea, methods, technology and approaches. Though it is mainly following the PIA application framework, for internal information fetching, in extraction method and filtering it looks for different and separate algorithms and approaches to make the whole system more accurate and robust.

11 Conclusion and Future Work

11.1. Conclusion

PIC fulfills all the design requirements in order to implement the system successfully. It has also solved all the research questions of the proposed research. Currently, the deployed system is successfully performing on the desktop. Its slideshow view can be displayed on any kind of device, especially from picture frame. PIC grabs information as an URL and processes it to create a note. Later, the node content is analyzed by using an external tool called Maui. Using Maui as a data extraction tool keeps the data private and secure. PIC is developed on drupal content management system, therefore for content analysis it has several modules for example-alchemy, kwanalysis, key word research etc. Maui has different measures to evaluate the performance of the keyword extraction and this features of Maui makes it more useful than drupal keyword extraction modules. As a filtering technique PIC is using drupal content-based module such as similar by terms. To filter out the similar nodes this module creates a common view by using a view parameter named contextual filter. This filtering parameter filters out all the tagged nodes inside the personal information center according to their node it. Therefore, our personal information center is a perfect solution to the need for a data aggregation service for the private cloud.

11.2. Future Work

Currently, PIC is able to auto-tag all the similar nodes content that have the same keyphrases or vocabulary terms and make an common view with the content of the similar nodes. We have already seen a common view with keyword "events" in the evaluation chapter but in the future we are planning to build an advanced knowledge based filtering system that can learn from the users behavior and suggests different services to the users. For example, after learning about the current location of the user PIC filtering system can suggest the weather information of that place to the user. Another example could be perhaps the user visited mountain several years ago with his/her family and may be the user has received an email where it has a keyphrase "mountain". Then the system will show the picture stored in the local folder related to that mountain tour or picture stored in the other private and public cloud storage such as dropbox, iCloud etc. Therefore, another future plan could be to add a private cloud storage platform such as ownCloud ¹ with the PIC so when the users receives any important email PIC can filter out emails using its knowledge based filtering system and save it to the different storage location using ownCloud platform according to the users' preference.

¹ <http://owncloud.org/>

12 List of Tables

8.1. Protocols used in PIC	45
8.2. Installed drupal 7 modules for PIC	45
9.1. Performance evaluation of keyword extraction	50

13 List of Figures

3.1. PIA system design[3]	12
3.2. Keyphrase Extraction using KEA[19]	14
6.1. Overall activity of system design	24
6.2. System design	25
6.3. Importer classes	26
6.4. Importer	27
6.5. Collaboration diagram of the extraction process	28
6.6. Class diagram of the extraction process	28
6.7. State Diagram	29
7.1. Deployment Diagram	31
8.1. Maui and Drupal 7 communication using drupal custom module	33
8.2. Activity diagram for PIC system implementation	34
8.3. Fetching information in PIC	35
8.4. Parsing information in PIC	36
8.5. Mapping for node processor	38
8.6. Pop up view of the Keyphrases extraction using Maui	38
8.7. Similar node in a common view	47
9.1. Similar node in a common view	51

14 Listings

8.1. Custom Maui keyword extraction module	39
8.2. Rules for autotaging content	42
8.3. Custom Rules Autotag module	43

15 Bibliography

- [1] S. S. Park, Y. S. Kim, and B. H. Kang, "Web information management system: Personalization and generalization.," in *ICWI*, p. 532, 2003.
- [2] G. Rossi, D. Schwabe, and R. Guimarães, *Designing personalized web applications*. ACM, 2001.
- [3] e. a. Albayrak, Sahin, *Agent technology for personalized information filtering: the PIA-system*. ACM symposium on Applied computing. ACM, 2005.
- [4] D. Kuropka and T. Serries, "Personal information agent," in *Pro. Informatik Jahrestagung*, pp. pp. 940–946., 2001.
- [5] W. Winiwarter, "Pea-a personal email assistant with evolutionary adaptation," *International Journal of Information Technology*, vol. 5, no. 1, 1999.
- [6] M. Balabanovi and Y. Shoham., *Fab: content-based, collaborative recommendation.* "Communications of the ACM 40.3 (1997): 66-72. Communications of the ACM, 1997.
- [7] H. Lieberman *et al.*, "Letizia: An agent that assists web browsing," *IJCAI (1)*, vol. 1995, pp. 924–929, 1995.
- [8] U. Shah, T. Finin, A. Joshi, R. S. Cost, and J. Matfield, "Information retrieval on the semantic web," in *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 461–468, ACM, 2002.
- [9] N. Pudota, A. Dattolo, A. Baruzzo, F. Ferrara, and C. Tasso, "Automatic keyphrase extraction and ontology mining for content-based tag recommendation," *International Journal of Intelligent Systems*, vol. 25, no. 12, pp. 1158–1186, 2010.
- [10] P. Turney, "Learning to extract keyphrases from text," 1999.
- [11] W. W. W. C. (W3C), *W3C Semantic Web Activity*. World Wide Web Consortium (W3C)., November 7 2011. Retrieved November 26,2011.
- [12] M. Wooldridge, *Agent-based software engineering*. IEEE Proceedings-software, 1997.
- [13] M. Wooldridgey and P. Ciancarini, *Agent-oriented software engineering: The state of the art.* "Agent-oriented software engineering. Springer Berlin Heidelberg, 2001.
- [14] T. Xia and Y. Chai, "An improvement to tf-idf: term distribution based term weight algorithm," *Journal of Software*, vol. 6, no. 3, pp. 413–420, 2011.
- [15] T. Miranda, M. Claypool, A. Gokhale, T. Mir, P. Murnikov, D. Netes, and M. Sartin, "Combining content-based and collaborative filters in an online newspaper," in *In Proceedings of ACM SIGIR Workshop on Recommender Systems*, Citeseer, 1999.

- [16] A. Moukas, "Information discovery and filtering using a multiagent evolving ecosystem," 1996.
- [17] B.-T. Zhang and Y.-W. Seo, "Personalized web-document filtering using reinforcement learning," *Applied Artificial Intelligence*, vol. 15, no. 7, pp. 665–685, 2001.
- [18] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "Kea: Practical automatic keyphrase extraction," in *Proceedings of the fourth ACM conference on Digital libraries*, pp. 254–255, ACM, 1999.
- [19] O. Medelyan and I. H. Witten, *Thesaurus-based index term extraction for agricultural documents*. Chapel Hill, NC, USA: ACM/IEEE-CS, June 11-15 2006. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries.
- [20] J. Braams, *Learning to extract keyphrases from text.*, vol. ERB-1057. Technical report NRC (ERB- 1057), February 1999.
- [21] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 01, pp. 157–169, 2004.
- [22] M. Dostál and K. Jezek, "Automatic keyphrase extraction based on nlp and statistical methods.," in *DATESO*, pp. 140–145, 2011.
- [23] L.-W. Lee and S.-M. Chen, "New methods for text categorization based on a new feature selection method and a new similarity measure between documents," in *Advances in Applied Artificial Intelligence*, pp. 1280–1289, Springer, 2006.
- [24] S. Doan and S. Horiguchi, "An efficient feature selection using multi-criteria in text categorization," in *Hybrid Intelligent Systems, 2004. HIS'04. Fourth International Conference on*, pp. 86–91, IEEE, 2004.

A Evaluation Matrix

	Requirements	Drupal	Joomla	Wordpress	Typo3
Input	Collecting /Fetching Information:				
	Public Sources :				
	RSS Feeds	x	x	x	x
	News Feeds	x	x	x	x
	Weather info	x	x	x	x
	Picture from Web Cam	x			
	Private Sources:				
	Emails	x	x	x	x
	Calendar	x			
	Social Network data (Optional)	x	x	x	x
Output	JSON (Preferred)	x			
	RSS	x			
	Atom	x			
Playlist View	N No. of playlists per user (N=1,2...N)	x			
	M No. of Inputs per Playlist (M=1,2...M)	x			
	X No of displayed items of a plugins per Playlist	x			
Visualization	HTML5 User Interface e.g Browser :	x			
	Render information from RSS Items e.g News Feeds	x			
	Render information from Email Items e.g Image				
	Render information from Email Items e.g Image				
Process					
	Integrate/aggregate data from Different sources	x			
	Refine/ optimize them in preferred way e.g ranking, prioritizing	x			
	Visualize the output in a User Interface e.g Browser	x			
Optional features	Image resize and modify	x	x	x	x
	Present all integrated information in Image	x			