A FRAMEWORK FOR AGGREGATING AND RETRIEVING RELEVANT INFORMATION ON-THE-FLY IN SUPPORT OF MAIZE PRODUCTION

PHILEMON NTHENGE KASYOKA

MASTER OF SCIENCE

(Computer Systems)

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

2014

A framework for aggregating and retrieving relevant information on-the-fly in support of maize production

Philemon Nthenge Kasyoka

A thesis submitted in partial fulfillment for the degree of Master of Science in Computer Systems in the Jomo Kenyatta University of Agriculture and Technology.

2014

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature:_____ Date:_____

Philemon Nthenge Kasyoka

This thesis has been submitted with our approval as University supervisors.

Signature:_____ Date:_____

Prof. Waweru Mwangi

JKUAT, Kenya

Signature:_____ Date:_____

Dr. Michael Kimwele

JKUAT, Kenya

DEDICATION

This work is dedicated to my wife Agnes Pendo and son Zion Mumo for giving me easy moment during my studies.

ACKNOWLEDGMENTS

I would like to thank the All Mighty God by whose will and grace this thesis has been a success. I would also like to thank my supervisors Professor Waweru Mwangi and Doctor Michael Kimwele for many insightful ideas in this study and for motivating and helpful guidance throughout this thesis. To all of you, may God the almighty shower you with his blessings.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

DFD	Data Flow Diagram
FAO	Food and Agricultural Organization of United Nations
HTML	Hyper-Text Markup Language
HTTP	Hyper-Text Transfer Protocol
ICT	Information and Communication Technology
IDF	Inverse document frequency
RSS	Really Simple Syndication
SPSS	Statistical Package for Social Sciences
TF	Term Frequency
TFIDF	Term Frequency Inverse Document Frequency
TREC	Text Retrieval Conference
URL	Uniform Resource Locator
XML	eXtensible Markup Language

ABSTRACT

Conventional web browsing requires users to open a web browser and look for relevant content from multiple different websites which can be a tedious task. Really Simple Syndication (RSS) provides a mechanism to aggregate content from different webs and push the content to users at scheduled intervals. RSS can be a solution to the tedious conventional web browsing.

Maize is the main staple food in Kenya and there is need to provide farmers with the relevant information to support maize farming and consequently improve production. Due to information overload, getting the right information from the internet has been very elusive for many people interested in maize information. One of the reasons has been due to lack of effective ways of aggregating and retrieving relevant maize information based on user preferences.

A framework for aggregating and retrieving relevant information in support of maize production was proposed as a solution. The framework makes use of Really Simple Syndication technology and retrieves relevant information through the use of Term Frequency-Inverse Document Frequency (TF-IDF). A new hybrid approach of using TF-IDF that integrates Term Proximity with TF-IDF was used for better performance. This approach is able to ensure maize farmers get relevant information to assist them in maize production.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Information Communication Technology has become a critical factor for driving growth and productivity in global economies. Farming is becoming a more time-critical and information-intense business.

According to (Ajani, 2014) ICT can directly support farmers' access to timely and relevant information, as well as facilitate the creation and sharing of knowledge among the farming community. A push towards higher productivity requires an information-based decision making agricultural system. Farmers must be able to get information at the right time and place with ease

Maize (also known as corn) is a cereal grain, believed to have been domesticated at least 7,000 years ago when it was grown in Central Mexico. According to FAO (2011) maize is the main staple food in Kenya, it accounts for about 40 percent of daily calories and has per capita consumption of 98 kilograms; this translates to between 30 and 34 million bags. There has been a fluctuating trend in maize production over the last decade, which threatens household food security and income sources. According to (De Silva & Ratnadiwakara, 2008) the cost of information from planting decision to selling at the wholesale market can make up to 11% of total production costs.

Over 85 percent of the rural population derives its livelihood from agriculture, most of who engage in maize production. In Kenya maize production accounts for roughly 20 percent of gross farm output from the small-scale farming sector (Jayne, *et al.,* 2001).

With great advancement in technology, availability and affordability of portable mobile devices that can access internet, more people have become regular seekers of online information. According to (Ndwiga et al., 2013) there is a need to make information on innovations available to maize producing farmers since they are literate enough to read and adopt the innovation to improve their production.

Finding relevant information online is not an easy task and there is a need for an efficient and effective online agricultural information system that will allow easy information retrieval by end users. Information provided by agricultural information systems is potentially useful to farmers and their communities, as well as various other stakeholders interested in the improvement of farmers' well-being.

In july 2006, there were over 1.6 million blog postings everyday. These numbers are staggering and suggest a significant shift in the nature of Web content from mostly static pages to continuously updated conversations.

Fannin and Chenault (2005) found that utilizing RSS (Really Simple Syndication) feeds for an agriculture news website led to increased awareness for the public and many website developers are increasingly using the RSS to publish content. According to (Brewington and Cybenko 2000) RSS technology is been used by publishers to publish news and article feeds and it can easily deliver up to-date posting. RSS can play a major role in the development of a framework that will allow users to get relevant and current information from the internet. RSS (Really Simple Syndication) is a powerful and simple web technology that makes it possible to easily access frequently updated content

on the Internet. Many websites have become more useful with the advent of RSS technology (Hendron, 2008).

According to (Isah, 2012) this technology is useful in many disciplines and has been used to support journalism, academic research, intelligence gathering, marketing and advertisement, communication within an organizations or professional groups as well as in media sharing. Feeds are designed for delivering web content updates, and are often referred to as RSS which is a defined standard based on XML (eXtensible Markup Language) and have been released in several versions which includes; RSS 0.90, RSS 0.91, RSS 0.92, RSS 1.0, RSS 2.0 and Atom.

Web feeds are never limited on the amount of content that can be published; they are able to fetch all topics that are periodically updated on a website. There is a need to make maize information more accessible using RSS technology as it is able to aggregate information from different sources and have it availed to users as soon as it is published.

There are numerous successful applications of RSS technology in practical use. One example is the use of RSS in libraries to manage large amounts of continuously changing, increasing, and/or updating information. RSS technology provides useful information to online library users with the newest items and most reserved books (Zeki, 2004).

RSS is a simple technology that does not have a mechanism to filter information it aggregates therefore there is a need to filter information delivered through the use of RSS technology to ensure readers get the most relevant content based on their

preferences. With greater technological development in the area of Information Retrieval, Term Frequency-Inverse Document Frequency (TF-IDF) has been adopted as an effective term weighting method for retrieval of relevant information in a vector space model. In a study conducted by Ramos (2003) TF-IDF was found to have a discriminatory power that allows retrieval engine to quickly find relevant documents.

A framework for aggregating and retrieving relevant information on-the-fly using TF-IDF and Term Proximity is been proposed as a solution. Through such as framework maize farmers are guaranteed to get the most relevant agricultural information, this will lead to effective access to information and more productive farming.

1.2 Statement of the problem

Despite the fact that a large population of Kenyans are farmers, the use of online Agricultural Information systems has been very minimal as importance of improved agricultural information in developing economies increases (Kizito, 2011).

Rapid increase in websites and blogs has led to an increase in information overload and it has become extremely difficult for farmers to locate current and relevant information to support maize production (Saravanan, 2010). A mechanism for aggregating, retrieving relevant information and ensuring users get information on maize farming as soon as it is published is necessary.

In this age of greater personalization (Kantor, 2007) there is a need to make agricultural content more accessible and allow it to be efficiently explored. This research has endeavoured to develop a framework for aggregating and retrieving relevant

information. The resulting framework puts more focus on improving relevance at the top-*k* documents returned by Term Frequency Inverse Document Frequency based on user information preferences.

1.3 Justification

ICT is already showing a great potential in the delivery of information to the agricultural sector in both developed and developing countries (Zijp, 1994). Through computers and mobile devices most people are constantly seeking information online. The use of RSS technology can improve delivery of relevant content to maize farmers ensuring they are not overwhelmed by large amounts of information.

The search for relevant information online for most farmers is difficult and consumes much of their time; the framework will help maize farmers find the most relevant information based on their preferences.

Maize is the main staple food among rural households in Kenya. However, there has been a fluctuating trend in maize production over the last decade, which threatens household food security and income sources. Relevant farming information will help farmers improve maize production.

Maize farmers need latest information to help them in their day to day farm decision making process and such a framework is of great benefit to anyone seeking information on Maize farming. The framework for aggregating and retrieving relevant maize information can helps individuals, relevant government agencies and agricultural organisations to effectively deliver relevant maize information to farmers.

1.4 Broad Objective

The overall objective of this research was to create a framework that will be able to provide relevant information to maize farmers in support of maize production..

1.5 Specific Objectives

i) To investigate the use of RSS Technology can be used to deliver aggregated agricultural information from websites.

ii) To find out factors influencing the use of internet based technology in accessing relevant maize information.

iii) To explore the use of Term Frequency-Inverse Document Frequency (TF-IDF) in the retrieval of relevant maize information.

iv) To design a framework for aggregating and retrieving relevant maize information.

v) To implement and test the framework for aggregating and retrieving relevant maize information.

1.6 Research Questions

i) How has will RSS technology be used to aggregate maize information from different websites?

ii) What factors affect the use of internet based technologies in accessing relevant maize information?

iii) How will Term Frequency-Inverse Document Frequency (TF-IDF) be used to retrieve relevant maize information?

iv) How will the framework for aggregating and retrieving relevant maize information be designed?

v) How will a framework for aggregating and retrieving relevant maize information be implemented and tested?

1.7 Scope of Study

The study examined the use of web technology in support of maize farming in Kenya. The essentials examined included RSS technology, TF-IDF and Term Proximity in the design and implementation of a framework for aggregating and retrieving relevant agricultural information.

1.8 Organization of the Thesis

This thesis is organized as follows: Chapter one presents information on background of the study, maize farming, problem statement, objectives, research questions and justification. Chapter two covers literature review on TF-IDF, Term Proximity and other related web technologies. The methodology used in the thesis is explained in chapter three. Results and analysis of the thesis is in chapter four while chapter five covers the conclusion and recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

This chapter reviews theories related to Information Retrieval, Term Frequency Inverse Document Frequency, term proximity approaches, maize farming, web technologies and presents gaps in the literature.

2.1 The RSS Technology

RSS is a web technology that works by having the website author maintain a list of notifications referred to as RSS Feed on their website. RSS technology is used to easily deliver up-to-date postings such as personal weblogs and news to subscribers (Brewington and Cybenko, 2000). People who are interested in finding out the latest headlines can check the RSS Feed. Websites create an RSS document and save it with a .xml extension. Creation of RSS is generally automated and is updated as soon as there is any new information. Many sites publish the date and time of the update in the title. A special computer program called RSS aggregator or RSS channel is used to automatically access the RSS feeds from different websites on your behalf and organize the results for you.

RSS is written in XML programming language for use by programs such as an RSS aggregator and not by a human. RSS is structured in channel and each channel consists of a <title> tag for the item, a <link> tag to a web page with the actual information and a <description> tag to describe the information. Sometimes this description is the full

information you want to read and sometimes it is just a summary. The RSS information is placed into a single file on a website in a manner similar to normal web pages.

Producing an RSS feed is very simple exercise, large news websites and most weblogs are maintained using special content management programs. When authors add their stories to the website using the program's "publish" facility .Those programs also updates the RSS feed XML file, adding an item referring to the new post, and removing less recent items. Blog creation tools like Blogger and LiveJournal among others automatically create feeds. Authors can also update XML files by hand, just as they update their website content.

2.2 How RSS works

The RSS system is used to publish articles and news over the internet in a very simple way:

a) First a publisher will need to identify the web page information they would like to publish on their website and on other websites through RSS.
b) An XML file that defines RSS feed is created, it holds URL, title and summary of each page to display.
c) A person who wishes to read the feed will need a RSS reader; the RSS file will be loaded from the publisher using a URL link that will display title and summary of the feed.
d) By clicking on a link the page containing the whole story or article will be displayed on an aggregator or the reader will be redirected to the publisher’s website.

2.3 XML Format of RSS Feed

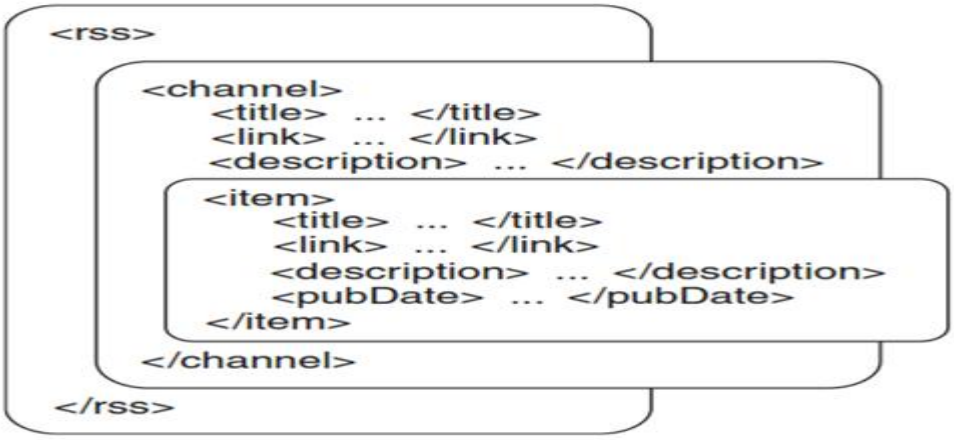

Figure 2.0 RSS Feed Structure

The figure above shows the basic structure of RSS 2.0 The subordinate to the <rss> element is the <channel> element that contains metadata about RSS feeds. <item> is an element that has information on postings. A channel may contain any number of <item>s. and they all have to be within the <channel> tag.

According to (Kim and Lee, 2005) RSS does not run under a push-based protocol but does under a pull-based protocol in which individual subscribers are responsible for collecting information from Web sites.

2.4 RSS aggregator

A web feed aggregator is an application that pulls relevant content from different RSS feeds to which a user has subscribed to, allowing the content to be read from within one location. According to Freeman (2009) a typical aggregator collects the feeds into folders based on the feed URL and the user can manually create some folders or group folders together. The main limitation is that as the number of subscriptions grows, the number of folders user has created and content also increases. This makes it difficult for

the user to select what to read or not to read. Users will have to quickly skim through each headline to determine if it is of interest to read or not to read.

RSS aggregators will automatically check a list of RSS feeds for new items every now and then to keep track of any changes made websites. They detect the additions and present them all together in a useful and compact manner. The role of RSS aggregator is important in web services and as the number of RSS feed users want to subscribe grows dynamically, the number of RSS feeds an RSS aggregator has to aggregate also grows. Postings dynamically will appear and disappear over time. (Kim and Lee, 2005) suggest that it will be important to have a good aggregation system that will efficiently aggregate posting generated and enables us to efficiently classify information from the postings and provide user with relevant content.

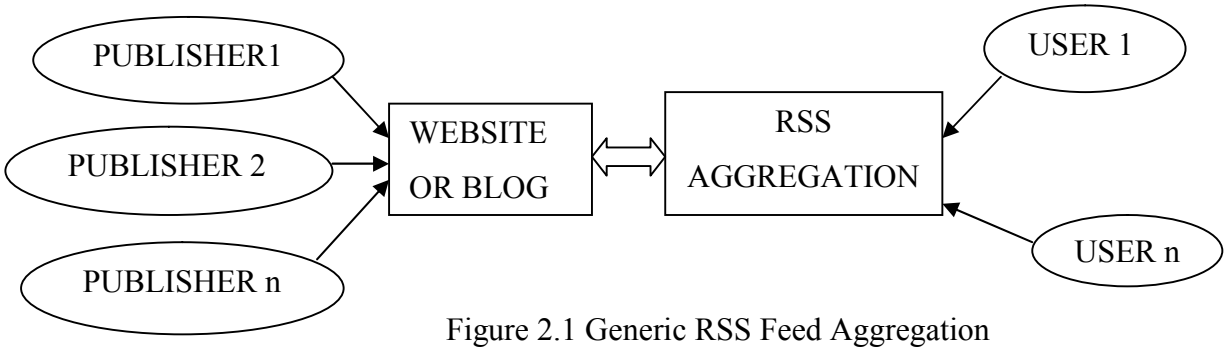

Figure 2.1 Generic RSS Feed Aggregation

2.5 Information Retrieval Models

According to (Gurkok, 2008) there are three main types of Information Retrieval models widely used in the field of information retrieval, namely:

2.5.1 Boolean Model

According to Indrawan (1998) The Boolean model is considered to be the simplest matching function in information retrieval.. It simply checks whether a keyword is present or absent in a document. This implies that term weights are assumed to be all in binary form of either 1 or 0. The use of 1 or 0 is not an effective approach in information retrieval. The query is formulated as a Boolean combination of keywords using operators and, or, and not.

2.5.2 Vector Space Model

Vector space is a statistical model which recognizes the disadvantages associated with the Boolean model. The vector space model represents both the documents and queries as a vector of Terms. It allows partial matching by assigning non-binary weights to index terms in queries and documents. These term weights are then used to compute the degree of similarity between documents and query. Salton and Buckley (1988) suggested various methods for assigning weights to index keywords.

2.5.3 Probabilistic Model

Unlike in the vector space model, in this model the document ranking is based on the probability of the relevance of documents and the query submitted by and end user. The probabilistic model starts with an initial guess of probabilistic description of the ideal set to retrieve the initial set of relevant document. By interacting with the user, the description of the ideal set is improved (Yates and Neto, 1999).

2.6 Information Retrieval and TF-IDF

According to Sanderson and Croft (2006), information retrieval (IR) system locates information that is relevant to a user's query. In Information Retrieval document is made up of terms which can be indexed, you can have a collection of hundreds of documents in your corpus, a term appearing in each document is useless as an index term because it does not tell anything about documents the user might be interested in, that is it does not act as a good discriminator while a term appearing in very few documents is quite useful and narrows the space of documents which might be of interest to the user.

The style of search used by both the electro-mechanical and computer-based IR systems was so-called Boolean retrieval. A query was a logical combination of terms (a synonym of word in IR literature), which resulted in a set of those documents that exactly matched the query. According to (Maron *et* al, 1959) Luhn (1958) proposed and they tested an alternative approach, where each document in the collection was assigned a score indicating its relevance to a given query. The documents were then sorted and those at the top ranks were returned to the user.

The researchers manually assigned keywords to a collection of 200 documents, weighting those assignments based on the importance of the keyword to the document. The scores assigned to the documents were based on a probabilistic approach. The researchers hand tested their ranked retrieval method, showing that it outperformed Boolean search on this test collection with 39 queries. Luhn (1958) suggested that the frequency of word occurrence in an article furnishes a useful measurement of word significance; his approach later became known as term frequency weighting.

This ranked retrieval approach to search was taken up by Information Retrival researchers, who over the following decades refined and revised the means by which documents were sorted in relation to a query. In the 1970s term frequency (*tf*) weights, was complemented with Spärck Jones's work on word occurrence across the documents of a collection. Her thesis on inverse document frequency (idf) introduced the idea that less common words tended to refer to more specific concepts, which were more important in retrieval(Jones,1972) and the idea of combining these two weights (tf-idf) was quickly adopted. Since then the use of TF-IDF has become popular in the field of Information Retrieval (Liu and Yang, 2012).

2.6.1 Term Frequency Inverse Document Frequency

According to (Soucy and Mineau, 2008) TF-IDF is the most common weighting method used to describe documents in the Vector Space Model. This weighting method is used in information Retrieval (IR) and it is composed of both the Term Frequency (TF) which is basically a count of the number of time a particular term appears in a text document. Inverse Document Frequency (IDF) form part of TF-IDF and is basically the count of all documents divided by the number of documents that contain the query term. According to (Zhang et al., 2005) IDF is calculated using equation (1).

$$IDF=\log(N/df) \qquad (1)$$

Term weights are a combination of Term Frequency (TF) and Inverse Document Frequency (IDF), considering relative document length as well as composition of the document collection (TF*IDF approach). Jones (2004) indicates that getting an effective

term weight score, Term Frequency is combined with IDF by simply multiplying the values as shown by equation (2).

$$w = tf_i \times \log(n/df_i) \quad (2)$$

Where $\mathbf{tf_i}$ is number of occurrences of a specific term in a document $\mathbf{n}$ is the total number of documents in the collection and $\mathbf{df_i}$ is the number of documents where the specific term appears at least once. The Log is basically used to normalize so those small documents are not treated unfairly against large documents.

2.6.2 The SMART System

The System for the Mechanical Analysis and Retrieval of Text (SMART) Information Retrieval System was developed at Cornell University in the 1960s. SMART is based on the vector space model, Salton and Buckley (1988) provide a number of variants of the TF-IDF weighting approach and present the SMART notation scheme where each weighting function is defined by triples of letters; the first one denotes the term frequency factor, the second one corresponds to the inverse document frequency function and the last one declares the normalization that is being applied.

The legacy of the SMART system belongs to the so-called SMART notation, a mnemonic scheme for denoting TF-IDF weighting variants in the vector space model are show in Table 2.0 where alphabets are used to denote the scheme used in combination.

Term frequency	Document frequency	Normalization
n (natural): $\mathrm{tf}_{t,d}$	n (no): 1	n (none): 1
l (logarithm): $1+\log(\mathrm{tf}_{t,d})$	t (idf): $\log\frac{N}{df_t}$	c (cosine): $\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented): $0.5+\frac{0.5\times\mathrm{tf}_{t,d}}{\max(\mathrm{tf}_{t,d})}$	p (prob idf): $\max\left(0,\log\frac{N-df_t}{df_t}\right)$	b (byte size): $1/CharLength^{\alpha}, \alpha<1$
b (boolean): $\begin{cases}1, & \text{if } \mathrm{tf}_{t,d}>0\\ 0, & \text{otherwise}\end{cases}$		
L (log average): $\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$		

Table2.0: SMART Mnemonic Scheme

The mnemonic for representing a combination of weights takes the form **ddd.qqq**, where the first three letters represents the term weighting of the document vector and the second three letters represents the term weighting for the query vector. A good example is **{ltc.ltn}** where Document={ltn} and Query={ltc}.

2.7 Web Technologies

The choice of appropriate technologies is an important success factor in the development of web applications. The web is supported by four main technologies:

i) Universal Resource Locator

Universal Resource Locator is a unique address at which page information can be found. It basically specifies the location of a resource. The web must know where to get a requested page, for this to be successful a unique name and protocol is required {protocol://site-name/page-name} a good example http://www.jkuat.ac.ke/admin.php where http is the protocol www.jkuat.ac.ke is the site-name and *admin.php* is the name of the page to access.

ii) Hypertext Transfer Protocol

HTTP is a text-based stateless protocol, controlling how resources such as HTML documents or images, are accessed. It's a client serve protocol designed to allow exchange of information over the web. HTTP defines the type of request a browser can request and the type of response a server can give through http user can retrieve a page from a remote server.

iii) Hypertext Mark-up Language

Hyper-Text Markup Language (HTML) was introduced in 1989 as the way to create documents on the World Wide Web**.** HTML uses tags to indicate what you want to include on the html document, it uses angular brackets (< and >) to separate metadata from basic text and defines a list of what can go into these brackets, such as *em* for emphasizing text, *tr* for table rows and *td* for representing tabular data.

iv) eXtensible Markup Language (XML)

XML has been designed for ease of implementation, and for interoperability with both SGML and HTML (W3C Working Draft, November 1996). Based on the W3C recommendation, XML (eXtensible Markup Language (W3C 1998)) has experienced truly triumphant progress with regard to its use and proliferation within and outside the Web. With its capability to define flexible data formats in the simplest ways and to exchange these formats on the Web, XML offers the prerequisite to homogenize heterogeneous environments. XML documents are characterized by two distinct properties: well-formedness and validity. While well-formedness is inherently anchored

in XML, validity can be ensured by the Document Type Definition (DTD) and XML schemas.

2.8 Stemming Algorithms

Stemming is a pre-processing step in Text Mining applications and very important in most of the Information Retrieval systems. The main purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. It is necessary to identify each word form with its base form, to do this a variety of stemming algorithms have been developed. Each algorithm attempts to convert the morphological variants of a word to its root form where the key terms of a query or document are represented by stems rather than by the original words.

The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form that is reduce the total number of distinct terms in a document or a query. Stemming algorithms can be classified in three groups: truncating methods, statistical methods, and mixed methods. Each of these groups has a typical way of finding the stems of the word variants.

2.8.1 Truncating Algorithms

The algorithms are related to removing the suffixes or prefixes (commonly known as affixes) of a word. The most basic stemmer was the Truncate (n) stemmer which truncated a word at the nth symbol i.e. keep n letters and remove the rest. In this method words shorter than n are not truncated. Examples Lovins Stemmer, Paice Stemmer and Dawson Porters Stemming Algorithm.

According to (Melucci and Orio, 2003). Porters stemming algorithm is as of now one of the most popular stemming methods proposed in 1980. Many modifications and enhancements have been done and suggested on the basic algorithm. It is based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes. It has five steps, and within each step, rules are applied until one of them passes the conditions. If a rule is accepted, the suffix is removed accordingly, and the next step is performed. The resultant stem at the end of the fifth step is returned. The rule looks like the following:

<condition> <suffix> → <new suffix>

For example, a rule (m>0) EED → EE means "if the word has at least one vowel and consonant plus EED ending, change the ending to EE". So "agreed" becomes "agree" while "feed" remains unchanged. This algorithm has about 60 rules and is very easy to comprehend. Porter designed a detailed framework of stemming which is known as 'Snowball'. The main purpose of the framework is to allow programmers to develop their own stemmers for other character sets or languages. Currently there are implementations for many Romance, Germanic, Uralic and Scandinavian languages as well as English, Russian and Turkish languages. Based on the stemming errors, Paice (1994) reached to a conclusion that the Porter stemmer produces less error rate.

2.8.2 Statistical Algorithms

These are stemmers that are based on statistical analysis and techniques. Most of them remove the affixes but after implementing some statistical procedure. An example of such stemmers are N-Gram Stemmer, HMM Stemmer which is based on the concept of

the Hidden Markov Model where transitions between states are ruled by probability functions (Melucci and Orio, 2003).

2.8.3 Inflectional and Derivational Algorithms

This is another approach to stemming and it involves both the inflectional as well as the derivational morphology analysis. The corpus should be very large to develop these types of stemmers and hence they are part of corpus base stemmers too. In case of inflectional the word variants are related to the language specific syntactic variations like plural, gender, case, etc whereas in derivational the word variants are related to the part-of-speech (POS) of a sentence where the word occurs. Krovetz (1993) presented a stemmer which is known to as the Krovetz stemmer.

2.9 Term Proximity in Information Retrieval

The distance between query terms in the document is referred to as term proximity (TP), according to (Tao and Zhai, 2007; Buttcher, 2006) information retrieval effectiveness can be greatly improved by integrating term proximity score into a retrieval model. Recent work by (Song et al, 2008) indicates that using flexible proximity terms within an information retrieval model results in improved retrieval effectiveness.

Proximity searching is one method that has been used to reduce the number of text document matches and to improve the relevance of the matched text documents by using term proximity to assist in ranking. Terms that are really far apart in a document are likely not to participate in any relationship, because it's hard for them to have any linguistic or semantic connection at all over that distance.

Intuitively, if query terms are near in a document we say the term proximity values are high and such a document should be more relevant to the query than another document in which query terms are far away from one another, if other factors are the same for the two documents. If a query term are "Information Management" a document that will have the two terms close to each other will be more relevant that a document that has the two terms spread far from each other on the document. Term proximity is more important in web search than in traditional information retrieval systems, due to the fact that there are usually millions of relevant results to a query (Zhu et al, 2007).

Term relationships in the document are important even without considering the structure of the terms in a query. Term Proximity helps maximize the chances that the query terms are related in the document and also maximize the chances that they are related in the way the user intended.

2.9.1 Term Proximity Measures

According to (Tao and Zhai, 2007) the following are two main approaches to term proximity:

a) Span based approach

This approach is used to measure term proximity based on the length of a text segment covering all the query terms on a document. There are two methods used in span-based approach. The first method is Span also referred to as the full cover. According to (Hawking and Thistlewaite, 1995) it is defined as the length of the shortest segment within a document that covers all query term occurrences, including repeated

occurrences. Example *d* = t1; t2; t1; t3; t5; t4; t2; t3; t4 Given *d* is a sample document where *t* represents different terms on a document, the span for terms t1 and t2 is 7

The second method is MinCover and it is defined as the length of the shortest segment in document that covers each query term at least once. Using the sample document *d* the minimum cover for t1 and t2 will be 2.

b) Pair-based/ Distance approach

In this particular approach the goal is to measure the proximity by aggregating pair-wise distances between query terms found on a text document. Using our sample document *d* the term proximity for {t1,t2,t3} will need the aggregation of distance between {t1,t2},{t1,t3} and {t2,t3}.

The pair-based approach has the following methods:

1) MinDist which is minimum distance defined as the smallest distance between all pairs of matching terms.
2) AveDist which is the average distance defined as the average distance value of all pairs of unique matched query terms.
3) MaxDist which is the maximum distance defines as the largest distance value of all pairs of unique matched query terms.

Research by (Tao and Zhai, 2007) shows that the MinDist performs much better than other measures while the span based approach would perform much better if normalization was applied to the query terms. Recent study by (Cummins and O'riodan, 2009) indicated that MinDist is highly correlated with relevance.

2.10 Maize Farming in Kenya

According to (FAO, 2011) maize farming is a staple in Kenya and over 75% of the local production is provided by small farmers. Production of maize in Kenya has not been enough to satisfy the market demand and as a result of low production maize has to be imported from other countries. Projections show that this shortfall will only increase in the future. The maize prices have also been increasing in recent years

Transition of maize to a major crop occurred in Kenya during World War 1, when the colonial government encouraged farmers to plant maize for the war effort. At the same time, a serious disease epidemic in the traditional food crop, millet, led to famine and stocks of millet seed were consumed rather than saved for planting. By providing farmers with seed of a late-maturing white maize variety, the colonial government sped the transition from millet to a maize-based food economy. After the war, the development of export markets encouraged maize production and by 1930s, maize was established as the dominant food crop in much of Kenya and Tanzania.

Maize accounts for about 40% of daily calories and per capita consumption is 98 kilograms. The poorest households spend 28% of the annual household income on maize purchase. Because of this importance, improvement in maize production will be crucial to solving Africa's food security problems and alleviating poverty. It is associated with household food security such that a low-income household is considered food insecure if it has no maize stock in store, regardless of other foods the household has at its disposal.

2.11 Use of Information Systems to Support Agriculture in Kenya

i) M-Kilimo

This system started in September 2009 and provides agricultural and horticultural information, advice and support.. The service primarily targets individual farmers and will also be accessible to agriculture extension facilities. In house agricultural experts answer registered farmers' queries in English or Swahili. In the event that an agricultural expert is unable to respond at once, the helpline agent contacts the second-line consultants and reverts to the farmer within 24hours(Brugger, 2011).

ii) Kenya Agricultural Information Network (KAINet)

According to (Chisenga, 2006) Kainet was initiated in April 2006 in response to demand from the national and international community to promote information exchange and access among stakeholders in the agricultural sector to support decision-making, to promote innovation in agriculture, and to improve livelihoods. KAINet addresses the national policy to build a Kenyan national agricultural science and technology information system to promote knowledge-sharing links between the national research system and extension and other rural service providers in Kenya, as well as international information systems.

iii) SMS Sokoni

According to (Muriithi e. al., 2010) SMS SOKONI is a system developed to enhance farming through the use of short message service, it was developed by Agricultural Commodity Exchange (KACE) in partnership with Safaricom mobile phone provider. Any farmer anywhere in the country can access updated and reliable market information on prices and commodity offers at an affordable rate using their mobile phones. So far,

the service is easy to use, reliable, convenient and affordable. The average monthly usage of this service increased from 1,273 in 2006 to 24,716 in 2008, demonstrating its subsequent usefulness and eagerness of farmers to explore the market information and linkage systems.

2.12 Related Work

Research conducted by (Mukai and Aono, 2005) analyzed web contents aggregated using RSS Technology, user profile was created from user browsing history and content was retrieved based on a comparison between user preferences and document content. They used Term Frequency (TF) to analyze web content. Its weakness was that the use of Term Frequency provides local weight of terms and not global weight. Ignoring the global weight of a term would not retrieve the most relevant text documents.

A similar framework was developed by Nagao (2008) that recommended RSS web content using weighted TF-IDF. He concentrated on the construction of RSS document considering the channel element and item element. Term weight of title was calculated using the average TF-IDF of the entire document where a user query term was found in the title.

(Gebre et al., 2013) Used uni-grams to improve the performance of TF-IDF in Native Language Identification while (Liu and Yang, 2012) proposed improving TF-IDF by introducing additional parameter class frequency denoted as CF that makes use of CHI square for feature selection. Research on adhoc retrieval was done by (Rasolofo and Savoy, 2003) where proximity function was incorporated in an information retrieval

model referred to as BM25 to improve the performance of short queries. A window of a size of five or less was used. They showed that marginal improvements on larger retrieval on a test collection referred to as TREC can be achieved.

A similar proximity function using bi-term was developed by (Buttcher *et al*, 2006). Their function permitted the use of arbitrary window size extracted from each scored document. There have been more researchers who have been successful in employing proximity into a number of keyword based retrieval functions. (Tao and Zhai, 2007) did a research on term proximity and were able to show significant improvements on the use of short queries in term proximity model.

A framework for incorporating information about the proximity between all query terms into a TF-IDF retrieval model was developed by (Cummins and O'Riordan, 2009), the approach used in this thesis calculates the proximity scores on top documents returned by the retrieval model by integrating both the pair-based and span-based approach. (Song et al, 2008) theorized an approach for span where they considered an ordered list of query terms on a document and their position, the approach used in this research identifies all spans found within a document and pick only the minimum span and calculates the proximity score for document.

Most of the research modifies the query and document by removing stop words, a good information retrieval system should be able to perform well with or without stop-word, the approach used in this research does not remove the stop-words. To find out whether the term proximity approach used with TF-IDF can improve retrieval effectiveness we will use TF-IDF as our baseline.

2.13 The Research Gaps

Different researchers discussed in the literature have used different techniques to try to improve access to aggregated information. Current agricultural information systems have not made significant attempts to improve access to relevant agricultural information; users have increasingly been overwhelmed by information overload or have not been able to find relevant information on such systems or websites. Maize is a staple food in Kenya and it is evident that there is a need to make information systems that support maize production more accessible.

A high-level motivation for use of TF-IDF is that it incorporates knowledge about the distribution of a term in all documents into the similarity measure for individual documents. According to (Wartena *et al.*, 2010) TF-IDF is a powerful weighting measure in IR however its implementation considers each query term individually and its implementation does not consider term proximity as a feature that might improve level of precision in Information Retrieval.

Majority of researchers who have implemented term proximity in IR models have included a sliding window with a maximum distance limit hence raising the question as to the effectiveness of not considering all occurrences of key terms in the text document. According to (Broschart, 2012) use of sliding window limits the number of documents whose scores are influenced by proximity scores.

2.14 Conceptual Framework

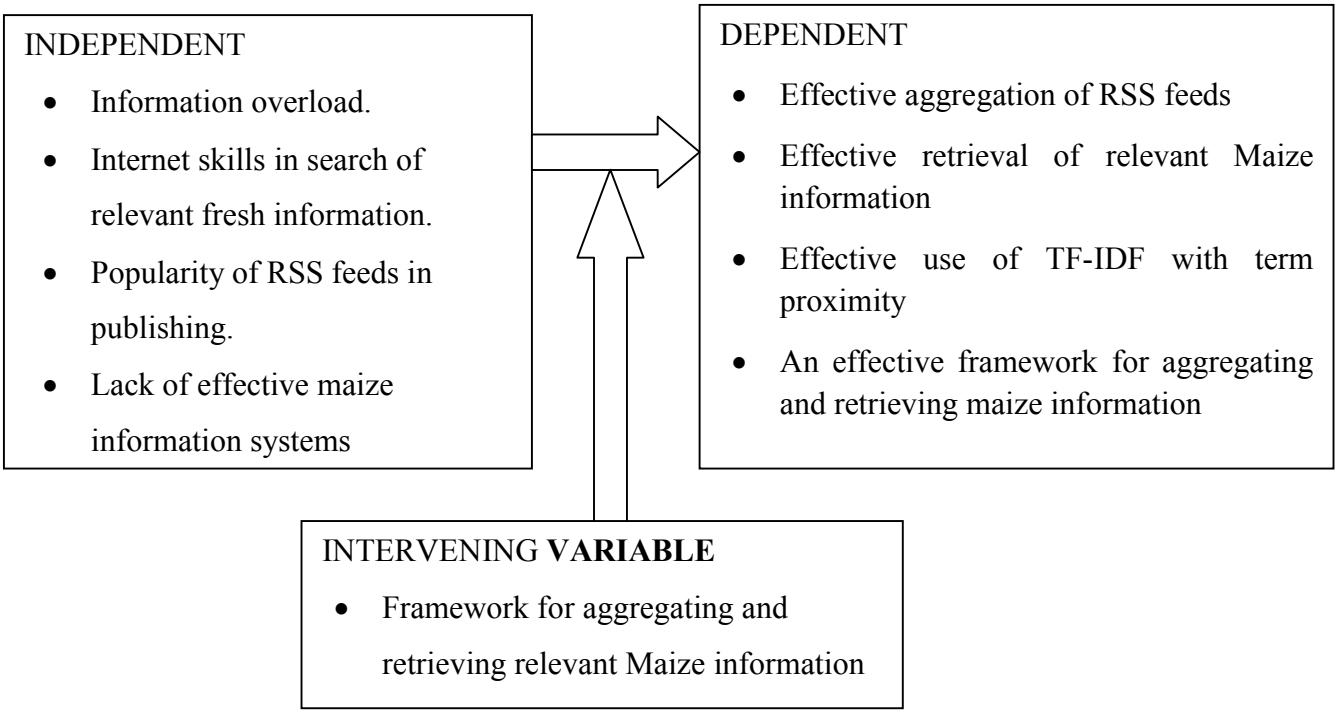

Figure 2.3 Conceptual Framework

2.14.1 RSS Aggregation process

The framework for aggregating and retrieving information makes use of RSS Technology to aggregate feeds from different agricultural websites and blogs. A simple RSS Feed XML format is shown below.

```
<rss version="2.0">
<channel>
<item> {Indicates start of article or story in RSS feed}
  <title> {Title of the article} </title>
  <description> {Describes the item} </description>
  <link> {Defines hyperlink to the item} </link>
  <date> {Date and time of publication}</date>
</item>
</channel>
</rss>
```

URL links for RSS feeds that provide maize information are stored in the database. Through aggregation process the framework picks RSS feed URL links from the database, through a fetch process and uses them to fetch RSS feeds from websites that have maize information, from the RSS feed the approach fetches only the ***title, description, link*** and ***date*** of the RSS feed and saves them in a database to form a text document Figure 2.4 shows the aggregation process.

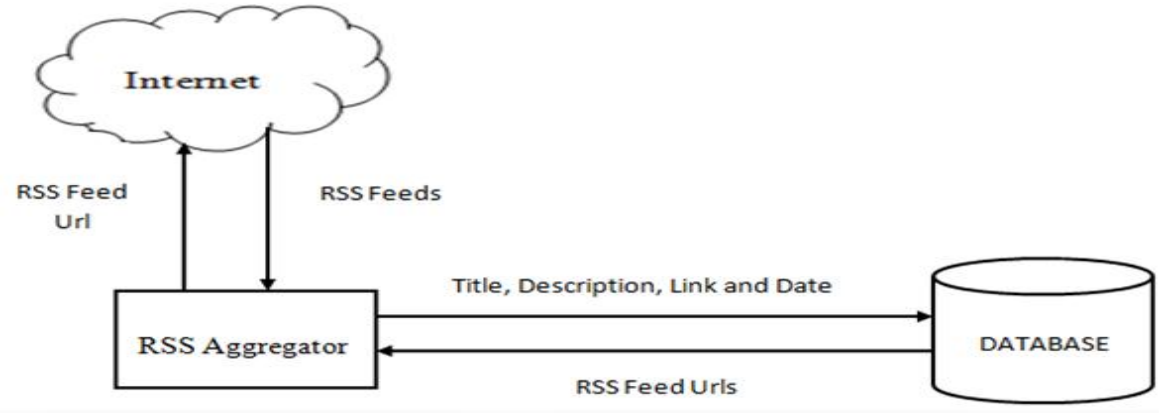

Figure 2.4 RSS aggregation Process

2.14.2 User Preference Profiling Process

User preferences are gathered when users register to use the web application that makes use of the framework. A new user is required to choose from a number of presented keywords such as **pests, diseases, fertilizer, *hybrid seed, harvesting*** and many more. Weights of the keywords are used to create a query vector which is measured against document vector using similarity metric. A user can change keywords later after registration if need be. User information collected during registration such as mobile phone number and email can be used to push information to users based on information relevance, this is done regularly according to user settings.

2.14.3 Stemming Process

Porters stemming algorithm which basically applies rules to remove suffix is used to convert query and document terms to their root form. Example **"harvesting"** can be stemmed to its root form **"harvest",** this helps improve precision of the information retrieval process.

2.14.4 Information Retrieval Process

Term Frequency Inverse Document Frequency (TF-IDF) is a powerful term weighting measure used in Information Retrieval and Inverse Document Frequency is calculated using equation (3) as normalized by Jones (2004).

$$idf = \log\left(\frac{dnum}{dfreq(T)}\right) + 1 \qquad (3)$$

Where *dnum* is the total number of documents in the corpus while *dfreq(T*) is the total number of documents where the term of interest appears atleast once. The SMART notation used for user query is {ltc} while the one used for document is {lnc}. TF-IDF is calculated using equation (4).

$$W_{t,d} = (1 + log(td)) \text{x } idf(\text{T}) \qquad (4)$$

Where $W_{t,d}$ is the overall TF-IDF while **1+*log(td)*** calculates the term frequency of a term of interest. The base of the logarithm used is immaterial since base 10, 2 or natural logarithm can be used. Query terms with high TF-IDF numbers imply a strong relationship with the document they appear in. A study conducted by (Buttcher et.al.,

2006). showed that query terms occurring close to each often result in a higher score hence retrieval effectiveness can be greatly improved by integrating term proximity score into a retrieval model.

According to Tao and Zhai,(2007) small distances between terms often imply strong semantic associations, thus we should reward cases where terms are really close to each other; however, when distances are large, the terms are presumably only loosely associated.

2.14.5 Calculating Similarity

A study conducted by (Peng & Guo,2013; Ahlgren *et al*., 2003) found cosine similarity performs better than Pearson's correlation. The cosine remains the best measure for the visualization of the vector space because this measure is defined in geometrical terms. Cosine correlation coefficient is used to calculate similarity between query vector and document vector as shown by equation (5) where similarity between two vectors is given by *y* (document vector) and *x* (query vector).

$$\mathrm{COS}(x,y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x^2_i} \cdot \sqrt{\sum_{i=1}^{n} y^2_i}} \qquad (5)$$

Where:

x_i is the TF-IDF weight of term *i* in the query.

y_i is the TF-IDF weight of term *i* in the document.

The cosine similarity provides a good metaphor and it is usually the cosine angle that

separates the two vectors *x* and y. If the similarity score is equal or close to zero it means the document and the query are not similar at all. If similarity score is from +0.8 to +1 it means that the query and the document are very similar.

Given query terms **{fertilizer, seeds, harvesting}** , a corpus of 2000 documents where document frequency is fertilizer=300, seeds=210 and harvesting=650 and sample document shown in figure 2.5 below.

The government of Kenya has promised to assist maize farmers with fertilizer and top quality seeds amid growing fears of fertilizer scarcity in the country.

Figure 2.5: Sample Document 1

The results of calculating similarity between the user query terms and the sample document 1 in figure 2.6 using cosine similarity are shown by Table 2.1.

Table 2.1 Sample Cosine calculation

Terms	**Query**						**Document**				**Product**
	tf	tf-tw	df	idf	wt	n'lize	tf	tf-tw	wt	n'lize	
fertilizer	1	1	300	0.82	0.82	0.59	2	1.30	1.30	0.79	0.47
seeds	1	1	210	0.98	0.98	0.72	1	1	1	0.61	0.44
harvesting	1	1	650	0.49	0.49	0.36	0	0	0	0	0
										Cosine dot product score	0.91

Query Length= $\sqrt{0.82^2 + 0.98^2 + 0.49^2}$ = 1.37 Doc Length = $\sqrt{1.30^2 + 1^2}$ = 1.64

The lengths are used to normalize the term weights (tw) as shown below

$^{0.82}/_{1.37}$=0.59, $^{0.98}/_{1.37}$=0.72, $^{0.49}/_{1.37}$=0.36 and $^{1.30}/_{1.64}$=0.47, $^{1}/_{1.61}$=0.44

Cosine dot product score is calculated as below

Score=(0.59x0.79)+(0.72x0.44)+(0.36x0)= 0.91

The document is very relevant to the user query since it has a cosine score of 0.91. When the 0.91score is added to the Term Proximity score the resulting value will be used for overall document ranking.

The[1] government[2] will[3] provide[4] subsidized[5] fertilizer[6] and[7] seeds[8] to[9] maize[10] farmers[11] this[12] will[13] ensure[14] they[15] get[16] better[17] maize[18] seed[19] harvest[20] this[21] year[22] and[23] improve[24] the[25] country's[26] food[27] security[28].

Figure 2.6: Sample Document 2

Poor[1] use[2] of[3] fertilizer[4] can[5] have[6] long[7] term[8] harmful[9] effects[10] to[11] soil[12] and[13] seeds[14], it[15] can[16] causes[17] seeds[18] not[19] to[20] germinate[21] or[22] lead[23] to[24] very[25] poor[26] maize[27] harvest[28].

Figure 2.7: Sample Document 3

The two sample documents shown in figure 2.6 and figure 2.7 above have the same term frequency for {fertilizer=1},{seeds=2} for user query terms.

Table 2.2 Query and Document vector Cosine calculation

Terms	**Query**						**Document**				**Product**
	tf	tf-tw	df	idf	wt	n'lize	tf	tf-tw	wt	n'lize	
fertilizer	1	1	290	0.84	0.84	0.52	1	1	1	0.52	0.27
seeds	1	1	210	0.98	0.98	0.61	2	1.3	1.3	0.68	0.41
harvesting	1	1	650	0.49	0.49	0.30	1	1	1	0.52	0.16
weeding	1	1	300	0.82	0.82	0.50	0	0	0	0	0
	Cosine dot product score										0.84

Given user query terms {fertilizer, seeds, harvesting and weeding} the query and document vector similarity is calculated giving a score of *SIM(Q,D)*= **0.84** as shown in table 2.2 above. Even though the two sample documents have the same similarity score compared to the user query terms relevance of the two documents might be different. In this research term proximity is implemented on the top-k documents returned by the information retrieval model to help distinguish such kind of documents.

2.14.6 Term Proximity Implementation

In a study conducted by (Cummins and O'riordan, 2009) they identified different pair-based approaches that have been used in term proximity, their research indicated that

minimum distance is highly correlated with relevance and therefore performs better than other methods, this corresponds with previous research work by (Tao and Zhai,2007).

In a study conducted by (Zhao and Yun, 2009) they found that use of span-based term proximity can lead better performance than other approaches of calculating proximity. In their research span-based proximity led to significant gain while used with ranking models. A proximity function that incorporates other proximity measures may outperform other proximity approaches (Cummins and O'riordan, 2009). In this research the approach used to calculate term proximity is a hybrid approach that will use both the span-based approach and pair-based approach to maximize on the strengths of each approach. Minimum proximity distance denoted as *min_dist* will be used for pair-based approach and minimum span denoted as *min_span* will be used for span-based approach.

The span-based approach used here is closely related to the one used by (Monz, 2004) where he defines the concept of matching span. However his definition was ambiguous, instead of defining the minimal matching span as the smallest segment that contains all query terms occurring in a document at least once, his span is checked to ensure that it does not contain another sub matching span within it. His approach does not effectively address the distribution of query term throughout the document. Research by (Tao and Zhai , 2007) shows that span-based approach works better when normalized by query terms. (Monz, 2004) does not normalize the minimal matching span with the number of query terms found within the matching span.

Our approach to span is slightly different, to get a *min_span* we will begin by counting the number of query term existing on the document, then scan the document for query

term starting from left to right, when all the query term have been found the segment of the document will form a span, we will begin checking for the next span from the next term until we find all the spans within the document. The next step will be to select the shortest span to become our *min_span* and normalize its distance with the number of unique query terms on the document. Example, given sample document *Doc*={A,J,D,C,P,T,D,X,Q,T} and Query={A,D,T} where alphabets stand for different terms on the document.

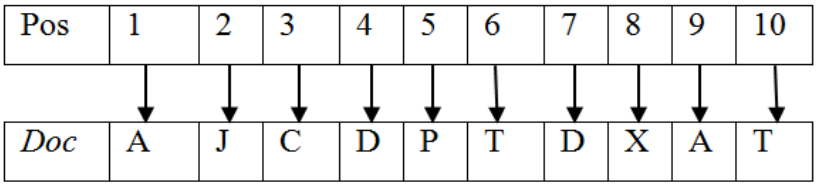

Figure 2.8: Term Position Mapping

Figure 2.8 shows how document terms are mapped against positions on a document, term position are counted from left to right starting from position 1 and so on. The spans for query {A,D,T} are {1,4, 6}=5 and {7,9,10}=3 ,since the order of query term is not relevant there the *min_span* is formed by query terms in positions 7,9 and 10.

When getting the distance in pair-based approach only the minimum distance between query terms will be considered, still using the example document and query for terms A, D will consider distance from position 7 and 9 and not position 1 and 4, for terms A,T we will look at position 9 and 10 and ignore position 1 and 6, for terms D,T we will consider position 4 and 6 and ignore positions 7 and 10.

The pair-based term proximity in this research is implemented using term scoring function that is closely related to the one defined by (Rasolofo and Savoy, 2003) where a feature that is proportional to the inverse square of the distance between each pair of queried terms is used. They limited query term consideration to a sliding window of size 5 and they calculated distance between all query terms occurrences on the document. Using such an approach the number of extracted pair dependencies grows exponentially with the number of query terms making longer queries impractical.

The scoring function used in this thesis differs from their work as the denominator in the term scoring function used to calculate the term pair index will only consider the minimum distance between term pairs $minDis(t_i,t_j)$ as shown by equation (6).

To ensure all the query terms found on the document have an equal opportunity to participate in the proximity scoring query terms will not be limited to a sliding window size.

$$TPi(t_i,t_j)= \frac{1}{minDis(t_i,t_j)^2} \qquad (6)$$

The *min_span* distance in the span-based proximity approach will be calculated for the shortest span on a document as shown by equation (7) then transformed to a span distance score using equation (8).

$$min\_cover= \left(\frac{ut_n^{pos} - ut_1^{pos}}{nt} \right) \qquad (7) \qquad Span= \frac{1}{(min\_cover)^2} \qquad (8)$$

Where: ut_n^{pos} is the last position of a unique query term noted on the text document, ut_1^{pos} is the first position of a unique query term occurrence noted on the document, *nt* is the number of all the unique query terms within the span.

Equation (9) shows how the *Span* score is integrated with the aggregated pair-based score to get the overall document term proximity score.

$$TP(Q,D)= \frac{\left(\sum_{i=1}^{n} TPi(t_i,t_i) \right) + Span}{n} \qquad (9)$$

$TPi(t_i,t_j)$ is the pair-based proximity score between two key terms, *Span* is the span-based proximity score of the document, *n* is the total number of all query terms on document. In previous work by (Tao and Zhai, 2007) to ensure a proper modelling of a term proximity score one of the conditions is that, the proximity score should decrease as the distance between query terms increases. The proximity score TP(Q,D) arrived at in this research satisfies that condition.

To ensure that the most relevant document are ranked high in users top-*k* a Relevance Score Value denoted as RSV is calculated and used as a ranking feature which is based on cosine score of query vector and document vector integrated with the overall term proximity score TP(Q,D) as shown by (10).

$$RSV=SIM(Q,D)+TP(Q,D) \qquad (10)$$

Where *SIM(Q,D)* is the cosine score. The term proximity score will be added to the results of the information retrieval model to improve the relevance of retrieved documents in user top-*k*.

2.14.7 Term Proximity for our Sample Documents

The sample document 2 and sample document 3 shown by figure 2.7 and figure 2.8 respectively have the same cosine similarity score of 0.8 therefore it would not be easy to rank the two documents in a way we can tell which is more relevant than the other. It is for such reason we need to implement the term proximity score on the top-k documents returned by information retrieval model.

Using our approach we will first calculate the term proximity score for the sample documents as shown below:

SAMPLE DOCUMENT 2

Span

Min_cover=(20-6)/3=4.7 span=$1/(4.7)^2$ =0.0453

Pair distance

TPi= $1/(8-7)^2 + 1/(20-19)^2 + 1/(20-6)^2 = 2.0051$

TP= (2.0051+0.0453)/4 =0.513

RSV=0.84+0.513 =1.3526

SAMPLE DOCUMENT 3

Span

Min_cover=(28-4)/3=8 span=$1/(8)^2$ =0.0156

Pair distance

TPi= $1/(14-4)^2 + 1/(28-18)^2 + 1/(28-4)^2 = 0.02174$

TP= (0.0156+0.02174)/4 =0.00934

RSV=0.84+0.00934 =0.84934

The calculations above shows that sample document 2 has a higher Term Proximity score of 0.513 hence a Relevance ranking score (RSV) of 1.353 compare to sample document 3 which has a Term Proximity score of 0.00934 and a RSV of 0.849. Sample document 2 will be ranked higher than sample document 3 meaning it is more relevant to the user query terms. The term proximity approach used in this research is in harmony with (Tao and Zhai, 2007) work which indicated that minimum distance is highly correlated with relevance.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter provides the methodology used in the research. It covers target population, research method, design, testing techniques, tools for analysis, technology for development and proposed architecture.

3.1 Research Design

In this study the researcher adopted a descriptive survey design. A descriptive survey research study was preferred since it has the dimension of investigating possible relationships between two or more variables. The descriptive survey design is ideal since it is concerned with making accurate assessment of the inference, distribution and relationship of the phenomenon. According to (Gay, 1981) descriptive research is a process of collecting data in order to answer questions concerning the current status of the subject in study.

An experiment was also conducted for purpose of evaluating the performance of TF-IDF and the proposed method.

3.2 Target Population and Sample

The population of interest in this study was maize farmers who use internet in search of agricultural information in Kenya.

The researcher adopted purposive technique where subjects with desired characteristics were identified using purposive sampling technique; the new identified subjects were

able to name other farmers with the required characteristics. This method was chosen because the target population was large and unknown. In this case the researcher was interested in maize farmers who use internet in search of agricultural information.

3.3 Instrumentation

The type of data used in the research study was primary data collected through a questionnaire. To ensure reliability of data collected a pre-test of the questionnaire was done to determine whether the respondents understood the questions correctly and where the questions did not seem clear enough, the necessary adjustments were made.

The questionnaire distributed to maize farmers contained both open ended questions as well as close ended questions. Questionnaire was chosen because of its simplicity of administration and high reliability as advocated by Babbie (1990). The items on the questionnaire were developed on the basis of the objectives of the study. The questionnaire contained three sections; Section A addressed background information, Section B sought information on factors that affect use of web based technology by maize farmers while Section C addressed technology preferences and use by maize farmers.

3.4 Reliability and Validity of the instrument

Reliability refers to a measure of the degree to which a research instrument yields consistent results or data after repeated trials. This type of reliability is referred to as Test-Retest. Test and retest simply put, is that you should get the same result on test 1 as you do test 2 when the two tests are administered after a time lapse. Retest involves two administration of the measurement instrument (Yin, 2003). The instrument was pre-

tested for their reliability where a number of 3 maize farmers were chosen for pre-test as a sample for pre-test should be small.

Alpha coefficient was used to test reliability of the instrument whereby a coefficient of 0.70 or more is acceptable. A high Cronbach alpha coefficient (0.7 and above) implies that the items correlate highly among themselves, that is, there is consistency among the items in measuring the concept of interest.

The sample for pre-test was also used to test data validity. The validation of the instrument was aimed at ensuring the instrument was measuring what they were intended to measure (Kathuri and Pals, 1993). The researcher also utilized experts in the agricultural field in order to ensure face and content validity of the instrument.

3.5 Data Collection Procedures

The instrument was administered by the researcher where the maize farmers were required to respond to questions asked by the researcher hence any clarification regarding the instrument was easily addressed.

3.6 Data Analysis

The study utilised first hand data which comes from the chosen respondents who answered the survey-questionnaires administered to them .In order to answer the research questions, the data collected needs to be thoroughly analyzed. Yin (2003) explains that every investigation should start with a general analytic strategy, allowing the researcher to decide what to analyze and why.

The questionnaires were edited for completeness and consistency before processing. Editing helped in detecting errors and omissions and which were corrected to ensure that maximum data quality standards were achieved. Data was then coded to enable responses to be grouped into categories. Coding involved assigning numbers so that the responses could be grouped into number of classes or categories. Data analysis was then carried out using the Statistical Package for Social Sciences (SPSS). The data collected was first subjected to descriptive statistics which included frequencies, percentages, means, and standard deviations. Inferential statistics was also very important for the study, in this case Pearson's moment correlation coefficient was used to determine the magnitude of relationship between two variables. A positive relationship means that an increase in one variable leads to an increase in another variable and vice versa.

3.7 System Design

UML which is a graphical modelling language will be used in the design of the system. Diagrams such as class diagrams, sequence diagrams, state diagrams, component diagrams and deployment diagrams will be used to describe major elements.

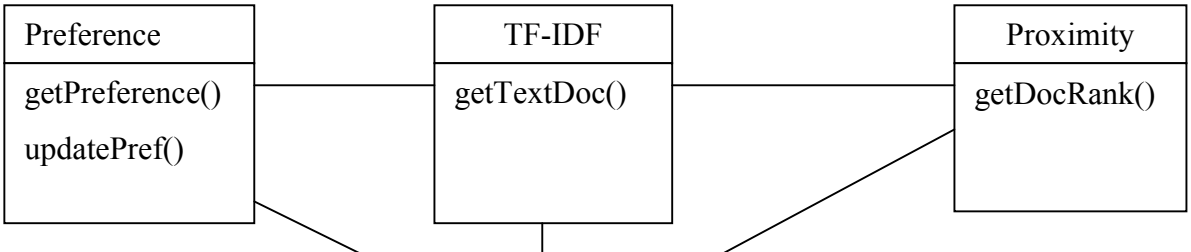

Figure 3.0: Class Diagram

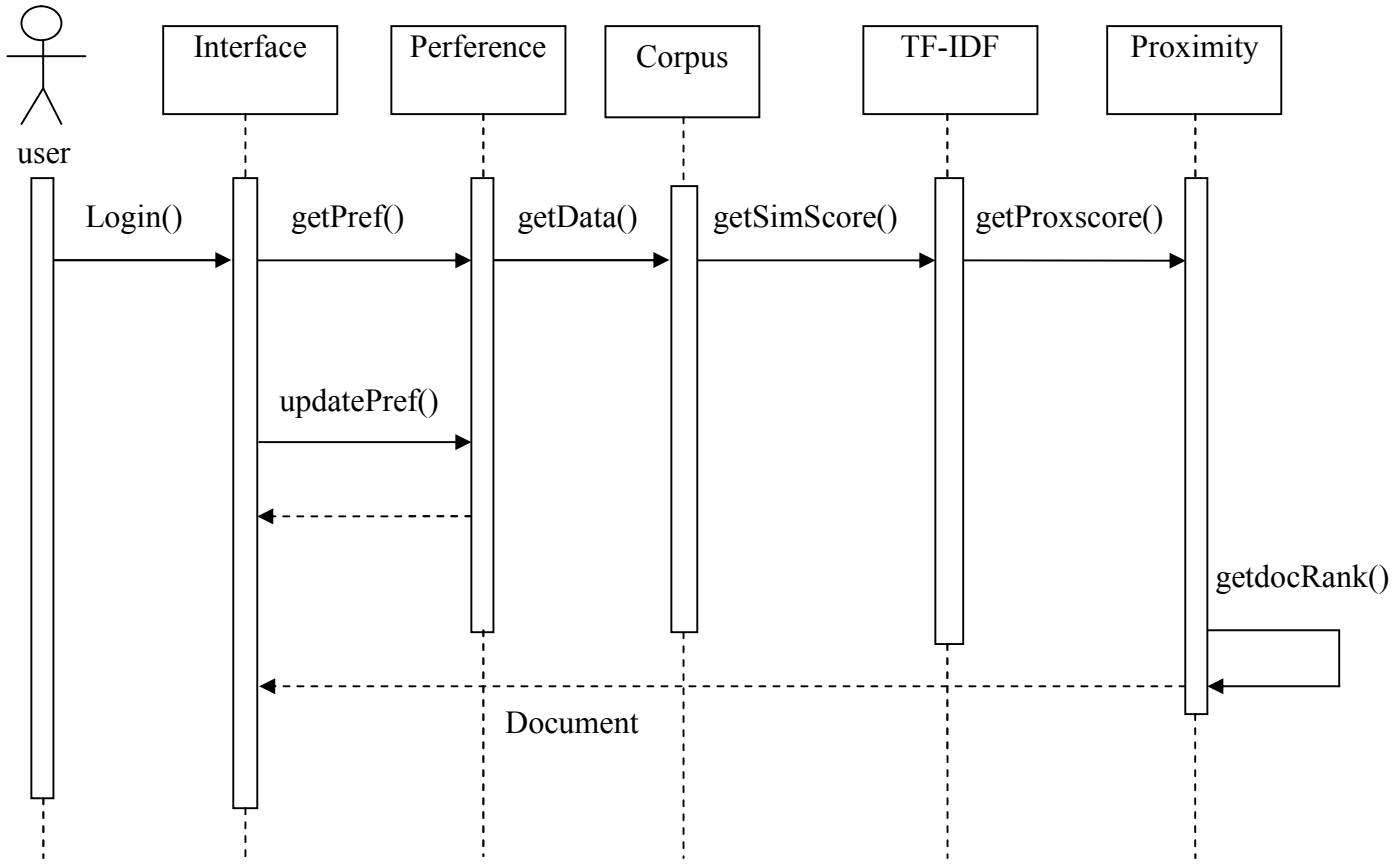

Figure 3.1: Sequence Diagram

3.8 Experiment and Test

For purpose of comparing performance of the proposed method, TF-IDF was used as the baseline. To determine the performance of the proposed method three groups each composed of two users were randomly selected as test user groups.

1) Set of user Queries

Set of queries were created for purpose of testing the framework, the queries were run on the baseline as well as the proposed method.

2) Relevance judgment

For each query created a set of relevant documents were identified, the identification process was done manually by use of different users who agreed on relevance of each text document to query presented to them. The relevance judgment list itself does not imply any ranking; it only contains the identification number of documents which judged relevant to the query.

3) Dataset

The experiment needed the use of agricultural dataset and since no standard agricultural dataset was found one was manually built which was composed of 232 RSS feeds collected from different agricultural websites.

The interaction of the test group in the information retrieval experiment is depicted by figure 3.2.

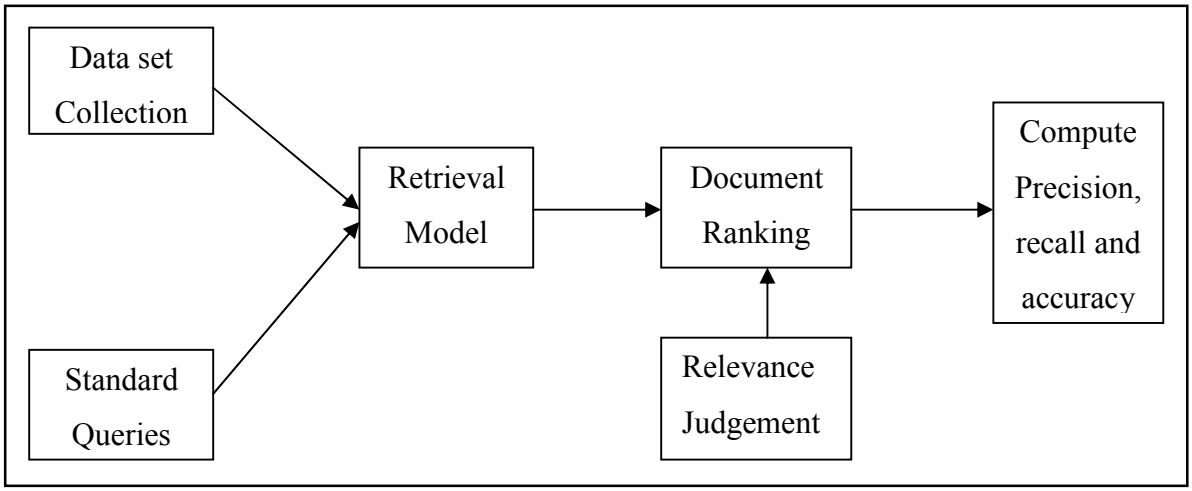

Figure 3.2: Experiment Model

Using the standard queries in a test collection queries were executed and data matching the queries was retrieved from the data set collection. Document retrieved were ranked based on relevance as calculated by the retrieval model where documents assumed to be most relevant to the query are ranked first. Human judgement is applied for relevance judgement on top 5 documents retrieved and on the top 10 documents retrieved. Precision which is percentage of relevant documents correctly retrieved by the system with respect to all documents retrieved is calculated based on relevance judgment using equation (11) while Recall which is the ratio of the number of documents retrieved to the total number of relevant documents in the corpus is calculated using equation (12)

$$\text{PRECISION} = \frac{\text{No. of Relevant documents retrieved}}{\text{No. of all Documents Retrieved}} \text{ X } 100 \qquad (11)$$

$$\text{RECALL} = \frac{\text{No. of Relevant documents retrieved}}{\text{No of all Relevant Document}} \times 100 \qquad (12)$$

$$\text{ACCURACY} = \frac{TP + TN}{(TP+FP+FN+TN)} \qquad (13)$$

Where:

TP is True Positive, TN True Negative, FP False Positive and FN False Negative.

3.9 Implementation

The framework was embedded in web application that can be used by farmers to access online information related to maize farming.

3.9.1 Relevant Information interface

Figure 3.3 below shows the main interface for a web application that uses the proposed framework to filter information. The left side displays the title of the most relevant information starting from the most relevant to the least relevant based on relevance score value while the right side displays the webpage of the title information selected on the left panel.

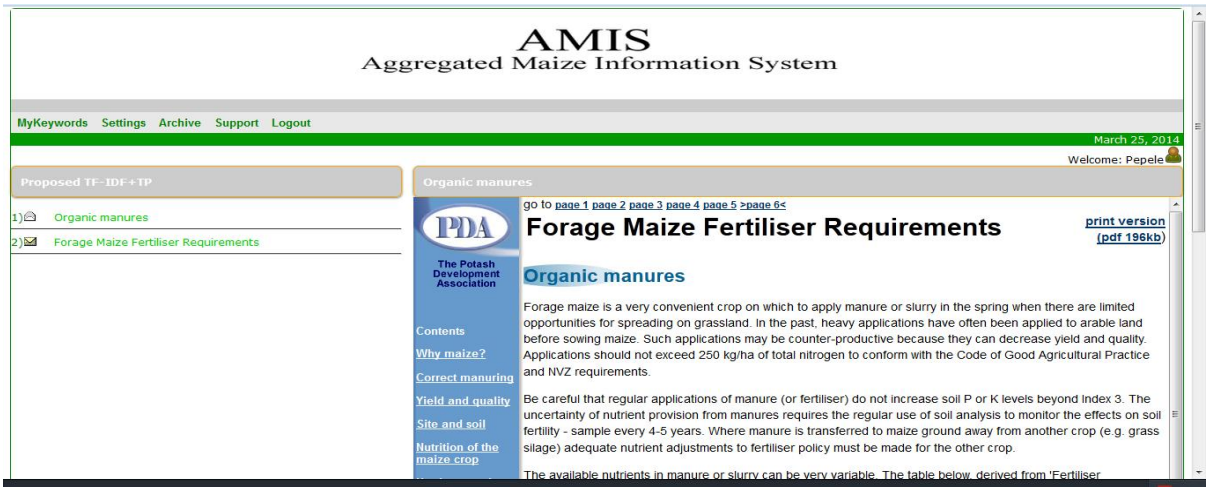

Figure 3.3: Screenshot of Framework implementation on web application

3.9.2 Keyword Panel

User preferences are composed of keywords that are provided in the keyword panel as shown in Figure 3.4 below. The panel shows the selected user preference keyword and provides more keywords that a user can select from to add to his list of preferences. One can add keyword by simply clicking on it and to remove keyword from preference list one needs to click on the keyword he/she wants to remove.

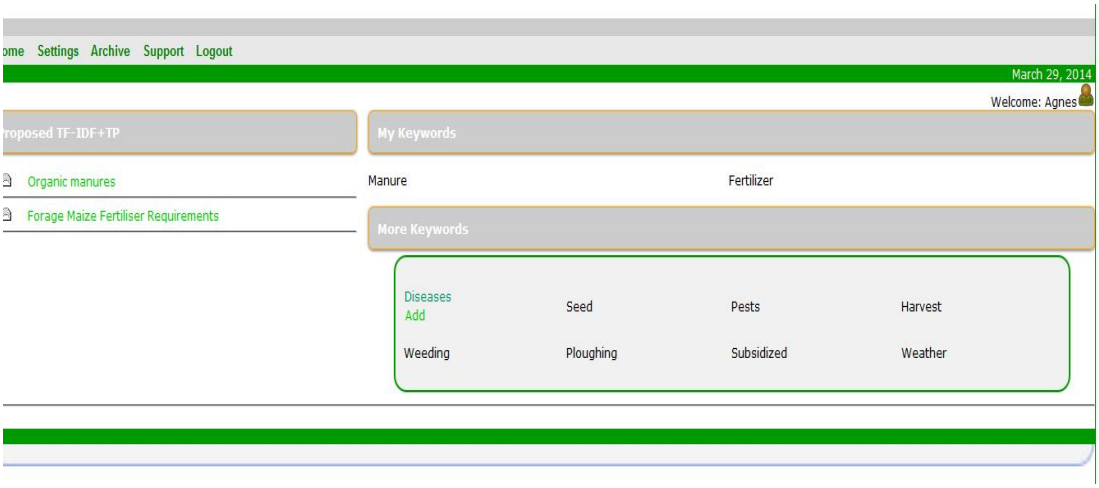

Figure 3.4: Screenshot of Preference Keywords

3.9.3 Settings Panel

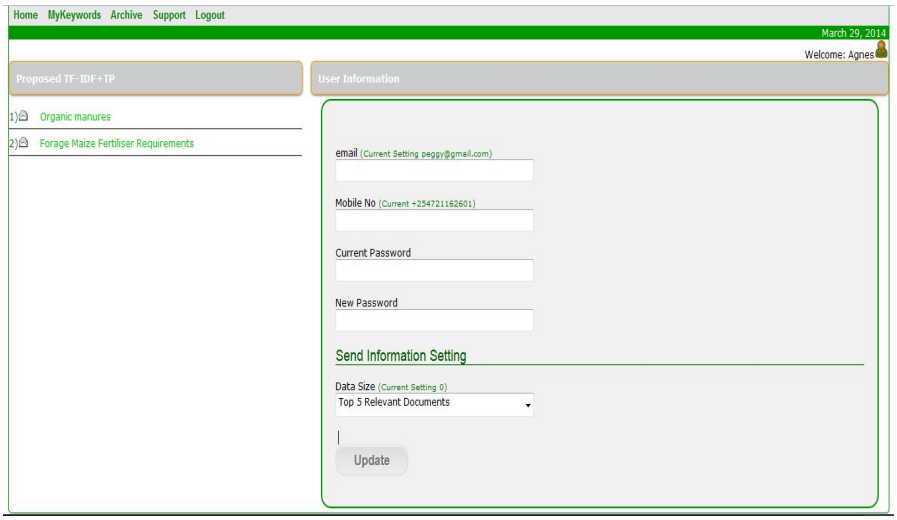

Figure 3.5: Screenshot of Settings Panel

Figure 3.5 above shows the settings panel which a user can use to edit profile information. Information such as email and mobile phone number can be used to send relevant information to users immediately it has been aggregated and ranked as relevant.

3.10 The Architecture of the Framework

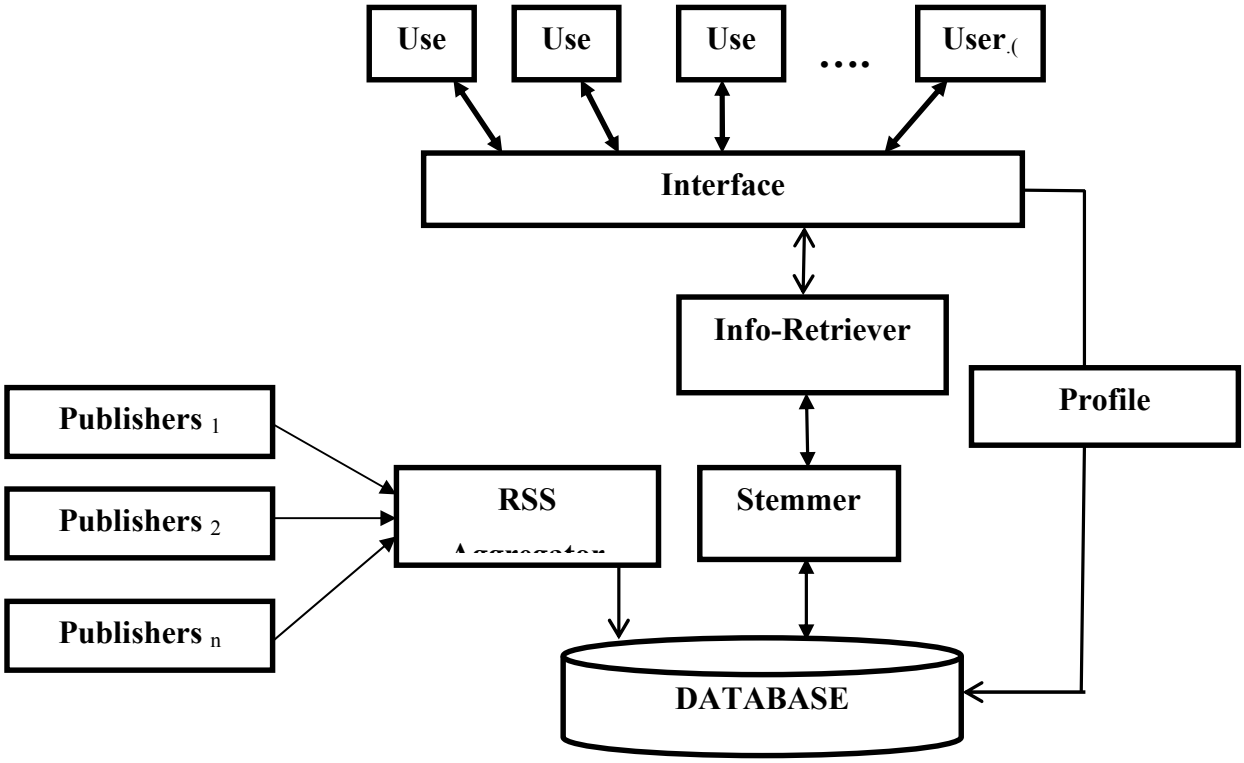

Figure 3.6: Architecture for Aggregating and Retrieving Framework

CHAPTER FOUR

RESEARCH RESULTS AND DISCUSSION

4.0 Introduction

This chapter presents the results of the analysis of data collected through the questionnaire and the experimental results of the proposed method and the baseline TF-IDF. The questionnaire results were interpreted and discussed in relation to the research questions raised in chapter one. The researcher interviewed maize farmers who use internet in search of agricultural information in Kenya. The responses were tabulated and subsequently presented by use of tables, bar charts and pie charts. The data was also used to determine descriptive statistics and inferential statistics. The findings were presented based on research questions as stated in chapter one.

4.1.0 Response rate

The questionnaire instrument was administered by the researcher and only 53 respondents were found matching the necessary requirements. 53 questionnaires were filled, one questionnaire got lost before data entry. Table 4.1.0 gives a summary of the information.

Table 4.1.0: response rate

	Frequency	Percentage
Responses	52	98%
Missing(Lost)	1	2%
Total issued	**53**	**100%**

4.1.1 Biographical information of the respondents.

Information on gender, age, academic qualification and computer/internet literacy was sought. 59.6% of the respondents were found to be male while 40.4% were female. This implies that most of the farmers are male and as such they are likely to seek for agricultural information online while few female farmers are likely to search for agricultural information online.

Table 4.1.1 Respondents Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	31	59.6	59.6	59.6
	Female	21	40.4	40.4	100.0
	Total	52	100.0	100.0	

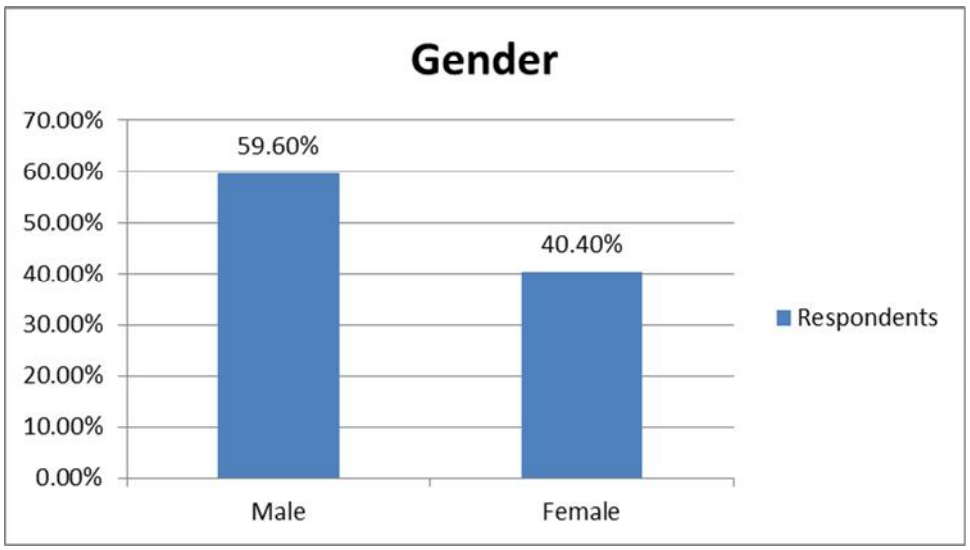

Figure 4.1.0 Gender of respondents

4.1.2 Age bracket

Information on the age of maize farmers was sought; table 4.1.2 shows the frequency distribution of the results. The study showed that most of the maize farmers using internet in search of farming information in Kenya are in the age bracket of 31yrs to 40yrs which accounts for 46.2% of the respondents. The results show that the number of respondent decreased as the age bracket increased signifying that the young generation is more likely to adopt technology in search of agricultural information. Table 4.1.2 summarizes the results.

Table 4.1.2 Age bracket

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18yrs-30yrs	8	15.4	15.4	15.4
	31yrs-40yrs	24	46.2	46.2	61.5
	41yrs-50yrs	17	32.7	32.7	94.2
	51yrs-60yrs	3	5.8	5.8	100.0
	Total	52	100.0	100.0	

4.1.3 Academic qualifications

Information on maize farmers' academic qualifications was sought. Figure 4.1.3 shows the summarized information. All the respondents were found to have some level of education ranging from diploma to postgraduate. Most of the respondents were found to have a bachelor's degree and this account for 38.5% of the population. The education level of the respondents in highly favours perceived usefulness and the basic skills necessary for use of web technology in search of agricultural information.

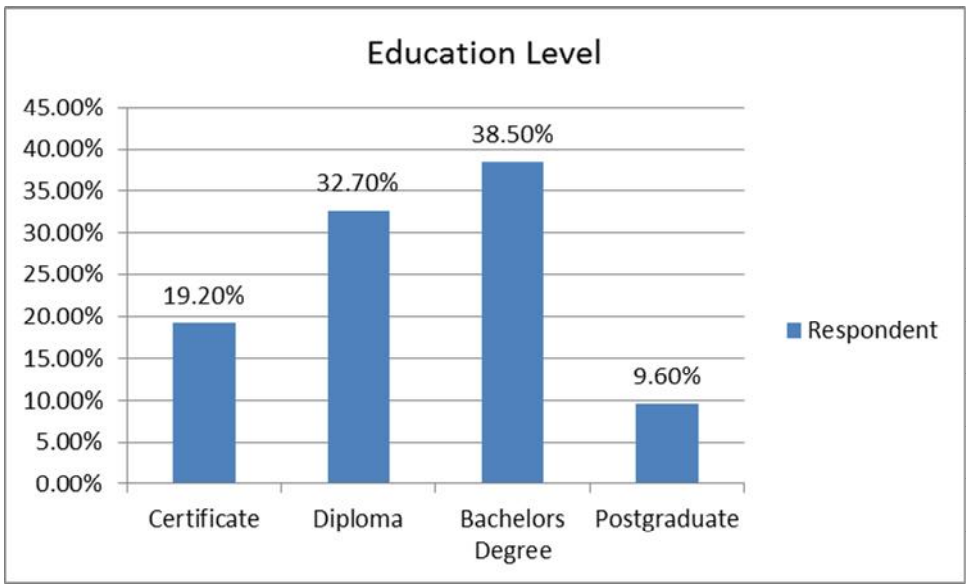

Figure 4.1.3 Maize farmers' education level

4.2.0 Factors Influencing use of Internet Based Technology

Information on the various factors influencing the use of internet based technology by maize farmers was sought under the following areas:

1. Farmers Income
2. Formal training on computer and internet technology
3. Trust of information delivered through internet technology
4. Perceived usefulness of internet technology in agriculture
5. Ease of access to internet infrastructure
6. Cost of internet connectivity
7. Ownership of devices able to access internet

4.2.1 Farmers Income

The researcher sought to determine the income level of the maize farmers. The results indicated that 32.7% of the respondents earn an income ranging from 26,000 to 50,000. The majority who make up 40.4% of the respondents earn a monthly income of between 51,000 and 75,000. 19.2% of the respondents earned between 76, 000 to 100, 000 while the rest who were 7.7% earned more than 100,000. The results are as presented by Table 4.2.1

Table 4.2.1 Farmers Monthly Income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	26,000 to 50,000	17	32.7	32.7	32.7
	51000 to 75000	21	40.4	40.4	73.1
	76000 to 100000	10	19.2	19.2	92.3
	over 100000	4	7.7	7.7	100.0
	Total	52	100.0	100.0	

4.2.2 Formal training and skills on computer and internet technology

Information regarding internet and computer literacy of the farmers was also sought it was found that majority of maize farmers who use internet had formal training on the use of computers as well as internet. Respondents were found to have internet and computer skill ranging from moderate to excellent favours use of web-based technology. 40.4 Percent of respondents indicated they had moderate internet and computer skills, 42.3% indicated they had good internet skills while 50% had good computer skills.

17.3% of the respondents had excellent internet skills while 9.6% had excellent computer skills as shown by figure 4.2.2. This gives an indication that the farmers training on use of computers most likely included internet skills.

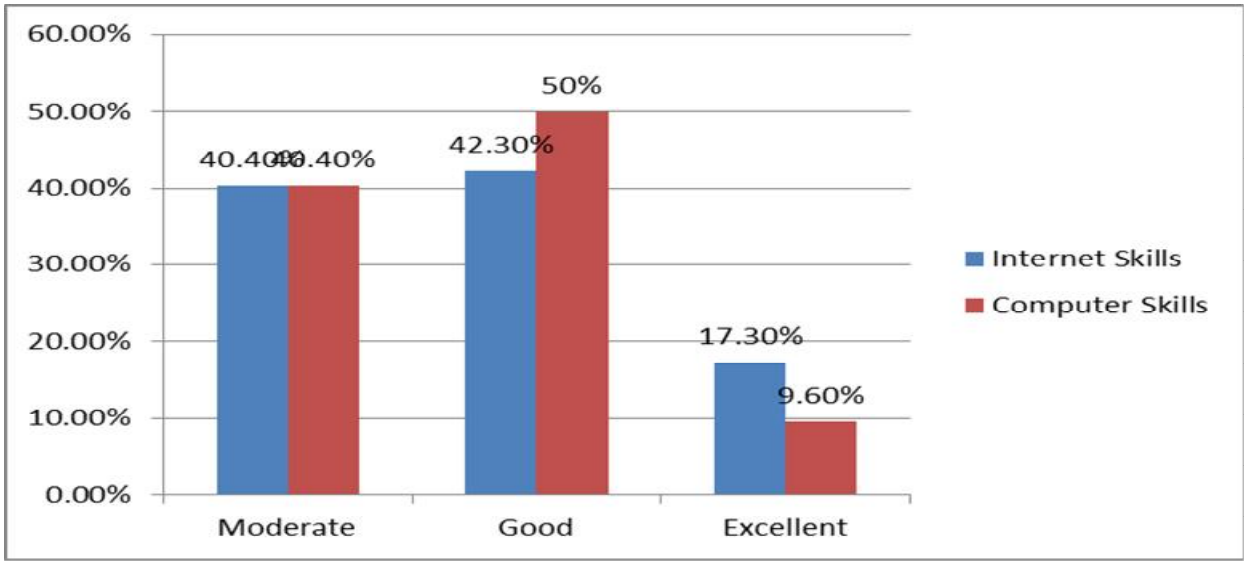

Figure 4.2.2: Computer and Internet Skills

4.2.3 Trust and perceived usefulness of internet technology.

Information regarding farmers' trust of information delivered through internet technology or found on agricultural websites was sought. 100% of the respondents trusted internet information where 57.7% of farmers agreed it was trustworthy while 42.3% strongly agreed. This has an implication that their trust of information found or delivered through internet based technology might have led to many farmers adopting and using online farming information.

Information regarding farmers' perceived usefulness of internet technology in delivering maize information was sought. 100% of the farmers interviewed believed that internet technology was useful in providing important farming information where 48.1% agreed and 51.9% strongly agreed. This highly contributes to the adoption of internet based technology in maize farming. Figure 4.2.3 summarizes the results.

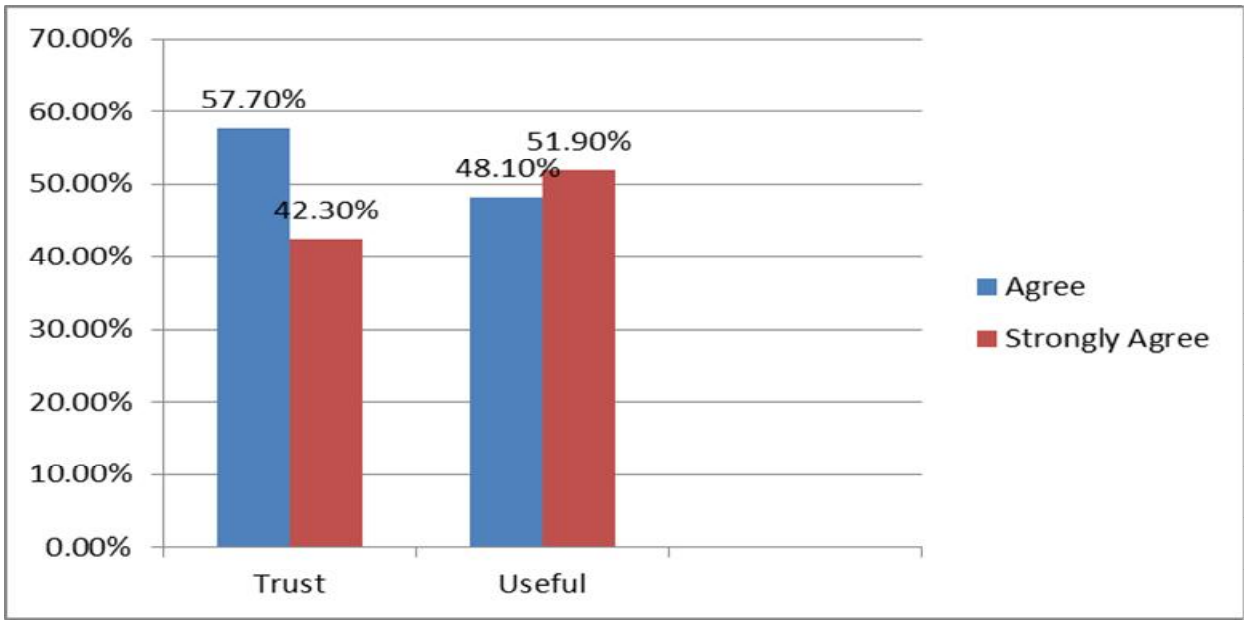

Figure 4.2.3 Trust and perceived usefulness of internet technology

4.2.4 Ease of access to internet infrastructure

Information on ease of access and availability of internet infrastructure was sought. Results show that a large percentage of farmers agree which account for 51.9% that they have easy access to internet, 36.5% strongly agreed, 7.7 % disagreed while the rest chose to remain neutral. Since a large percentage had access this might be contributing

to the fact that more farmers now are using internet to search for farming information and others willing to adopt web technology can easily do so. Table 4.2.4 gives a summary of the results.

Table 4.2.4 Ease of Access to Internet

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Disagree	4	7.7	7.7	7.7
	Neutral	2	3.8	3.8	11.5
	Agree	27	51.9	51.9	63.5
	Strongly Agree	19	36.5	36.5	100.0
	Total	52	100.0	100.0	

4.2.5 Device ownership

The researcher sought information concerning device ownership, the results show that all respondents who are 100% in number own web enabled mobile phones, 13.5% of then own laptops, 7.7% own desktops while only 1.9% own IPad/tablet. This clearly tells us that the majority have more than one device they can use to access information online and this favours the used to internet technology.

Table 4.2.5.1 Measures of Dispersion

	Monthly income	Computer Skills	Internet Skills	Trust of Information	Perceived Usefulness	Internet Infrastructure Access	Cost of Internet
N Valid	52	52	52	52	52	52	36
Missing	0	0	0	0	0	0	16
Mean	3.02	2.69	2.77	4.42	4.52	4.17	1.67
Std. Error of mean	.127	.087	.101	.069	.070	.116	.120
Median	3.00	3.00	3.00	4.00	5.00	4.00	2.00
Mode	3	3	3	4	5	4	1
Std. Deviation	.918	.643	.731	.499	.505	.834	.717
Variance	.843	.413	.534	.249	.255	.695	.514
Skewness	.594	.382	.392	.321	-.079	-1.185	.602
Std. Error of Skewness	.330	.330	.330	.330	.330	.330	.393
Kurtosis	-.416	-.644	-1.013	-1.975	-2.075	1.469	-.796
Std. Error of Kurtosis	.650	.650	.650	.650	.650	.650	.768
Range	3	2	2	1	1	3	2
Minimun	2	2	2	4	4	2	1
Maximum	5	4	4	5	5	5	3
Sum	157	140	144	230	235	217	60

Table 4.2.5.1 shows measure of dispersion and distribution it is clear that most of the factors highly favor the use of internet based technology this is shown by the mean, median, skewness and kurtosis. The mean and median of the different factors considered is not too different meaning that the distribution is symmetric this is confirmed by small

value of skewness. Kurtosis is a measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is zero while skewness which is a measure of the asymmetry of a distribution where the normal distribution is symmetric and has a skewness value of zero. The kurtosis for most the factors is close to zero and the value of skewness for the most of the factors presented is less than twice the standard error and is taken to indicate that the measure is within the symmetry therefore our the results of the factors presented here favour use of internet based technology in agriculture by maize farmers.

4.3.0 Farming Information

Farming information was sought from maize farmers in Kenya where the following categories of information were collected:

1. Farming Experience
2. Size of Maize farm
3. Type of farming
4. Type of land preparation used
5. Type of Maize variety planted
6. Use of chemicals in planting (fertilizer, pesticide and Herbicide)
7. Weather forecasting information
8. Maize yield and storage

4.3.1 Maize Farmers' experience

Information on maize farming experience was sought from the respondents. The research findings indicate farmers with less than 5 years of experience were 23.1%. 50% of the respondents had farming experience of between 5 years and 10 years .Respondents with 11 years to 15 years of experience were 21.2% while 5.8% of the respondents had experience ranging from 16 to 20 years.

Table 4.3.1 Maize Farmers' experience

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than 5yrs	12	23.1	23.1	23.1
	5yrs-10yrs	26	50.0	50.0	73.1
	11yrs-15yrs	11	21.2	21.2	94.2
	16yrs-20yrs	3	5.8	5.8	100.0
	Total	52	100.0	100.0	

4.3.2 Size of Maize farm

Information regarding the size of maize farm was also sought and the researcher established that most the respondents plant maize crop in farm lands whose size less than 5 acres, this accounts for 32.7%. The results clearly show that the number of respondents decreases as the size of acreage increases. In earlier findings we established that majority of maize farmers using internet were young farmers this tells us they might not be farming on large farms and this can have adverse effects on countries maize yield.

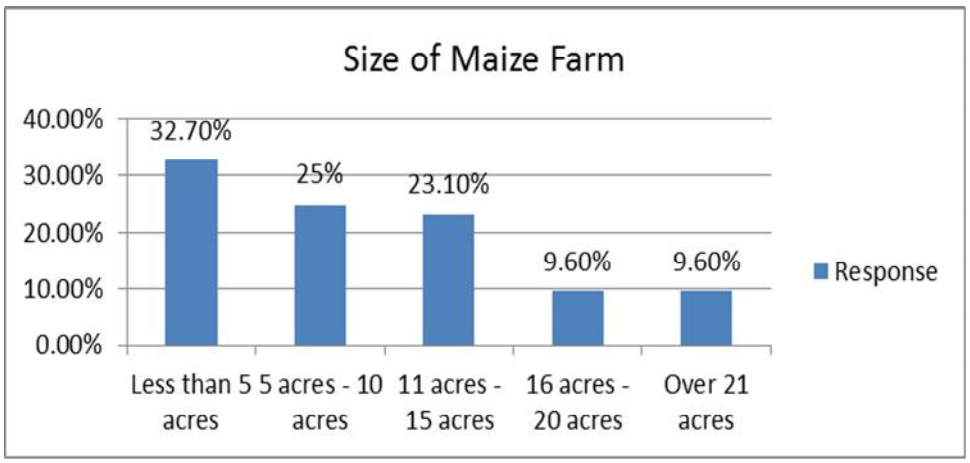

Figure 4.3.2 Size of Maize Farm

4.3.3 Type of farming

The researcher sought information about the type of farming most maize farmers are engaged in and it was established that 78.8% farmers who are the majority combine both subsistence farming and commercial farming. 17.3% practice subsistence farming while a small percentage of 3.8% grow maize for commercial purpose. This might explain the need for more farmers to seek more information to improve their maize production. Table 4.3.3 gives a summary the results.

Table 4.3.3 Type of farming

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Subsistence	9	17.3	17.3	17.3
	Commercial	2	3.8	3.8	21.2
	All Above	41	78.8	78.8	100.0
	Total	52	100.0	100.0	

4.3.4 Transportation Problems

For those farmers who planted maize for commercial purpose the researcher sought to know if they experienced difficulties with transportation of their maize produce to market places. 53.8% of farmers indicated they experienced problem. Through a follow up open question majority of the problems cited by the respondents were mostly poor road infrastructure and high cost of transportation of produce to market place and lack of information on where they can hire transport when required. The results are as presented by Table 4.3.4

Table 4.3.4 Maize Transportation Problem

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	YES	28	53.8	65.1	65.1
	NO	15	28.8	34.9	100.0
	Total	43	82.7	100.0	
Missing	System	9	17.3		
Total		52	100.0		

4.3.5 Preferred Land preparation method

Information regarding preferred land preparation method was sought and it was found that 46.2% of farmers prefer the use Tractor in land preparation this might be because a tractor is mechanical and much more efficient and faster than the use of animals or

manual labour. 26.9% of the respondents preferred to combine the use of Animal, Tractor and Manual labour in land preparation. Table 4.3.5 summarizes the findings.

Table 4.3.5 Preferred Land Preparation Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Animal	2	3.8	3.8	3.8
	Tractor	24	46.2	46.2	50.0
	Manual labour	12	23.1	23.1	73.1
	All above	14	26.9	26.9	100.0
	Total	52	100.0	100.0	

4.3.6 Type of Maize Seed Planted

The researcher sought information about the type of maize seeds planted by farmers. It was established that 80.8% of the respondents who are the majority plant hybrid seed on their farms. This might be because hybrid seeds germinate faster and are resistant to disease and harsh climatic conditions compared to the local maize seed variety. The hybrid variety used by most farmers was H6213 and OPV seed Maize TD1. Only a small group account for 19.6% of the respondents plant local variety which might be due to cost of the maize seed or difficulty in accessing the hybrid seed. Table 4.3.6 summarizes the results.

Table 4.3.6 Type of Maize Seeds Planted

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Local	10	19.2	19.2	19.2
	Hybrid	42	80.8	80.8	100.0
	Total	52	100.0	100.0	

4.3.7 Use of Fertilizer, Pesticide and Herbicide

Information regarding Frequency of usage of fertilizer, pesticides and herbicides was sought. 26.9 % of the respondents very often use fertilizer, 38.5% use fertilizer often and 34.6% use it rarely on their farms. A large percentage of maize farmers which accounts for 42.3% rarely use pesticide this could be as a result of not having a lot of pesticides on farms those who use pesticide very often account for 21.2% while 51.9 % who are the majority of maize rarely use herbicides of their maize farm. Figure 4.3.7 summarizes the findings.

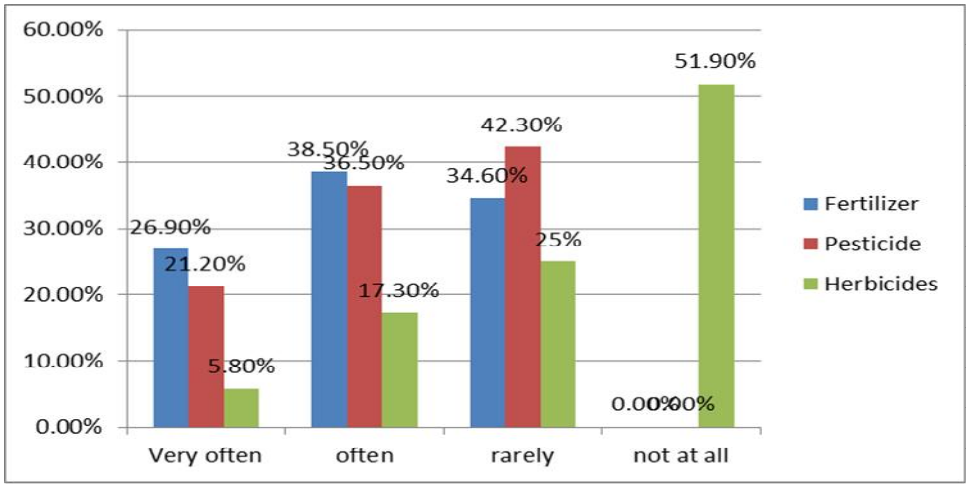

Figure 4.3.7 Use of Fertilizer, Pesticide and Herbicide

4.3.8 Maize Yield

Information regarding maize produce in previous season was sought. The results show that 25.5% which is the largest population of the respondents get a maize yield of 11 to 20 bags which is not a high yield. This might be as a result of majority of farmers interviewed farming on less than 5 acres size farms. Table 4.3.8.summarizes the findings.

Table 4.3.8 Maize Yield

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 11 Bags	11	21.2	21.2	21.2
	11 to 20 Bags	13	25.0	25.0	46.2
	21 to 30 Bags	6	11.5	11.5	57.7
	31 to 40 Bags	11	21.2	21.2	78.8
	> 40 Bags	11	21.2	21.2	100.0
	Total	52	100.0	100.0	

4.3.9 Timely information through Internet

The researcher sought to determine if maize farmers were finding the information delivered through internet more timely than their other sources of maize information. Most farmers accounting for 80.8% indicated they strongly agree that internet information was more timely compared to other sources. This might have contributed to most farmers' consistent use of internet technology.

4.4.0 Technology

Information on technology was sought from farmers using internet in search of agricultural information in Kenya and the following categories of information were collected:

1. Devices preferably used to access information
2. Current and Preferable means of receiving agricultural information
3. Frequency of access to online agricultural information
4. User-friendliness of agricultural website/blogs
5. Access to enough agricultural information

4.4.1 Devices preferably used to access information

The researcher sought information about the most preferable device used to access agricultural information by maize farmers. A greater percentage of maize farmers who account for 44.3% of the respondents would rather use their mobile devices to access online information. This might as a result of availability of affordable mobile devices that are web enabled and cheap internet connectivity provided by most of the mobile service providers. 36.5% of farmer would prefer the use of laptops this might be as a result of availability and cost of laptops going down while desktop are increasingly becoming out-dated. Table 4.4.1 summarizes the results.

Table 4.4.1 Preferred Devices for access of Internet Information

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Desktop	8	15.4	15.4	15.4
	Laptop	19	36.5	36.5	51.9
	Mobile	23	44.3	44.3	96.2
	ipad/Tablet	2	3.8	3.8	100.0
	Total	52	100.0	100.0	

4.4.2 Current and preferred means of receiving maize information

Information about farmers' current and preferred mean of receiving maize information was sought, the analysis shows that 50% access information directly on websites and 42.3% through email and another 7.7% access it through SMS. Preference on means of information access was sought and results shows that most farmers who account for 53.8% would prefer to access maize information through emails which is an increase from 11.5% compared with what they currently use. The percentage of farmers who prefer to use SMS was 25% which is a major increase from 7.7% of those who currently receive information through SMS. The percentage of those who accessed information on directly on website has dropped from 50% to 21.2% this shows that farmers are more willing to access information through technology that can provide information "on-the-fly". Table 4.4.2 and Table 4.4.2.1 summarize the results.

Table 4.4.2 Current Medium of Receiving Agricultural Information

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	SMS	4	7.7	7.7	7.7
	Email	22	42.3	42.3	50.0
	Website	26	50.0	50.0	100.0
	Total	52	100.0	100.0	

Table 4.4.2.1 Preferred Medium of Receiving Agricultural Information

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	SMS	13	25.0	25.0	25.0
	Email	28	53.8	53.8	78.8
	Web Application	11	21.2	21.2	100.0
	Total	52	100.0	100.0	

4.4.3 Frequency of access to online Maize Information

The researcher sought information about farmers frequency of access to online maize information, it was realized that majority of farmers accessed agricultural information 11 to 20 times in a month which accounts for 51.9%. Frequency is average which could indicate they might also be accessing information on other issues or the access of maize information might be seasonal.

Table 4.4.3: Access Frequency to Online Maize Information

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than 5	2	3.8	3.8	3.8
	6 to 10	20	38.5	38.5	42.3
	11 to 20	27	51.9	51.9	94.2
	21 to 30	2	3.8	3.8	98.1
	Over 30	1	1.9	1.9	100.0
	Total	52	100.0	100.0	

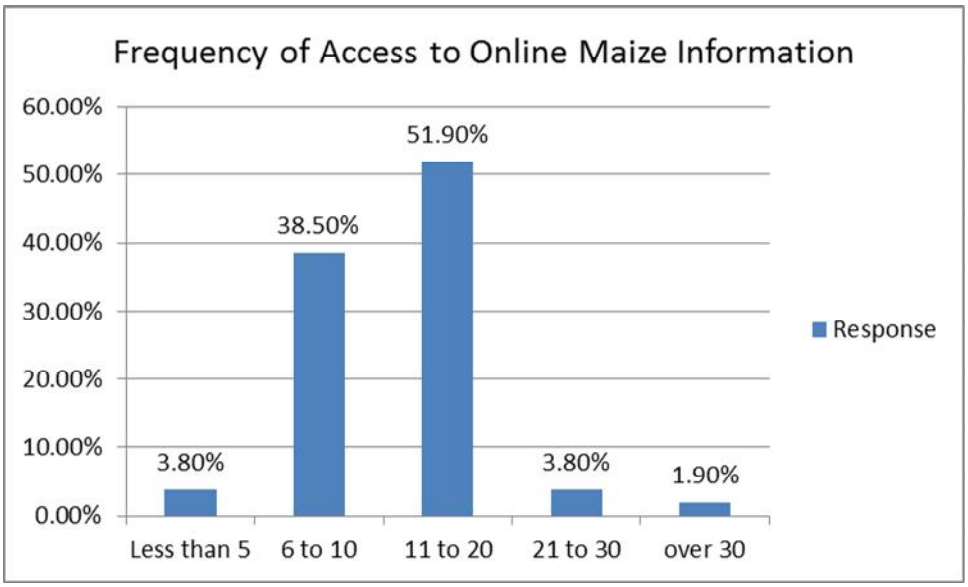

Figure 4.4.3 Access frequency to online agricultural information.

4.4.4 User-friendliness of agricultural website/blogs

Information on user friendliness of agricultural website or blogs farmers' access information from was sought. Even though majority of respondent indicated they had

good computing and internet skills as shown in section 4.2.2, majority of farmers who account for 55.8% did not agree that most of the websites were user-friendly while 17.3% were neutral. 26.9% of the respondents were in agreement that the agricultural websites were user-friendly. This is a clear indication that most websites are not user friendly; access of information from different sources might have also contributed to difficulties in user of the websites. There is a need to ensure websites providing maize information are user-friendly.

Table 4.4.4: User-friendliness of agricultural website/blogs

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Disagree	9	17.3	17.3	17.3
Disagree	20	38.5	38.5	55.8
Neutral	9	17.3	17.3	73.1
Agree	14	26.9	26.9	100.0
Total	52	100.0	100.0	

4.4.5 Access to adequate agricultural information

Availability of adequate maize information on websites or blog was sought. The results show that a great number of farmers who account for 80.8% did not find adequate information on the website they visited online while 11.5% were undecided. Only 7.7% of the respondents agree that they get enough maize information. This could be contributed by the fact that the maize information they are looking for is on different

websites and blogs aggregating to one web application may positively help improve access to adequate information. Table 4.4.5 gives a summary of the results.

Table 4.4.5: Access to adequate Maize information.

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Disagree	12	23.1	23.1	23.1
Disagree	30	57.7	57.7	80.8
Neutral	6	11.5	11.5	92.3
Agree	4	7.7	7.7	100.0
Total	52	100.0	100.0	

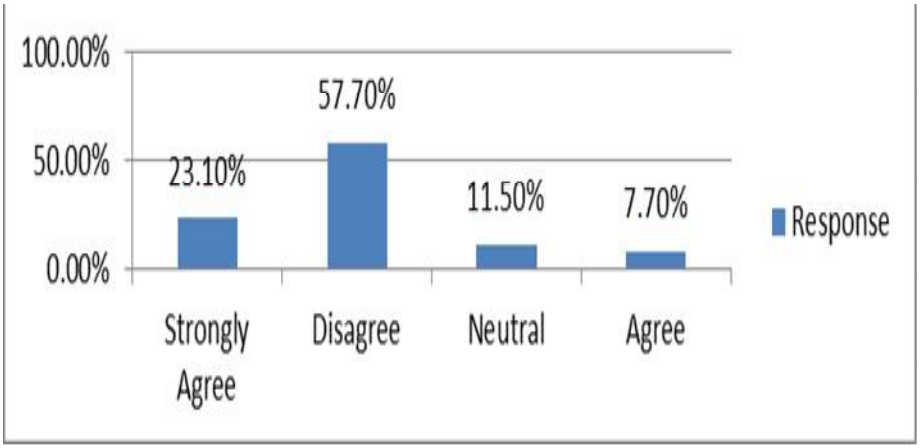

Figure 4.4.5 Access to Adequate Maize Information

4.4.6 Access to Relevant Agricultural information

The researcher sought information on relevance of maize information found on agricultural websites or blogs to their search queries. The results show that a great

number of farmers who account for 82.7% did not find information completely relevant to their search adequate, 1.9% were undecided while 15.4% agreed they were able to find information they sought relevant to their search queries. This might be as a result of lack effective information filtering method especially when a website has information on other crops besides maize crop or the fact that the maize information sought is scattered on different websites or blog. I would be helpful if information was aggregated in one central location. Table 4.4.6 gives a summary of the results.

Table 4.4.6 Access to Relevant Maize Information

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly Disagree	12	23.1	23.1	23.1
Disagree	31	59.6	59.6	82.7
Neutral	1	1.9	1.9	84.6
Agree	4	15.4	15.4	100.0
Total	52	100.0	100.0	

4.5 Test Experiment Results of the Framework

Queries created were executed on the dataset using the baseline method and the proposed method where precision, recall and accuracy was calculated for top 5 and top 10 documents retrieved using the two methods. Table 1.0 shows the results.

Table 4.5.0: Precision, Recall and Accuracy at Top 5 and 10 Documents

	RSS Title and Description					
METHOD	Precision @ 5	Precision @ 10	Recall@5	Recall@10	Accuracy @5	Accuracy @10
TF-IDF	0.69	0.54	0.33	0.26	0.73	0.69
Proposed	0.78	0.69	0.65	0.44	0.86	0.81

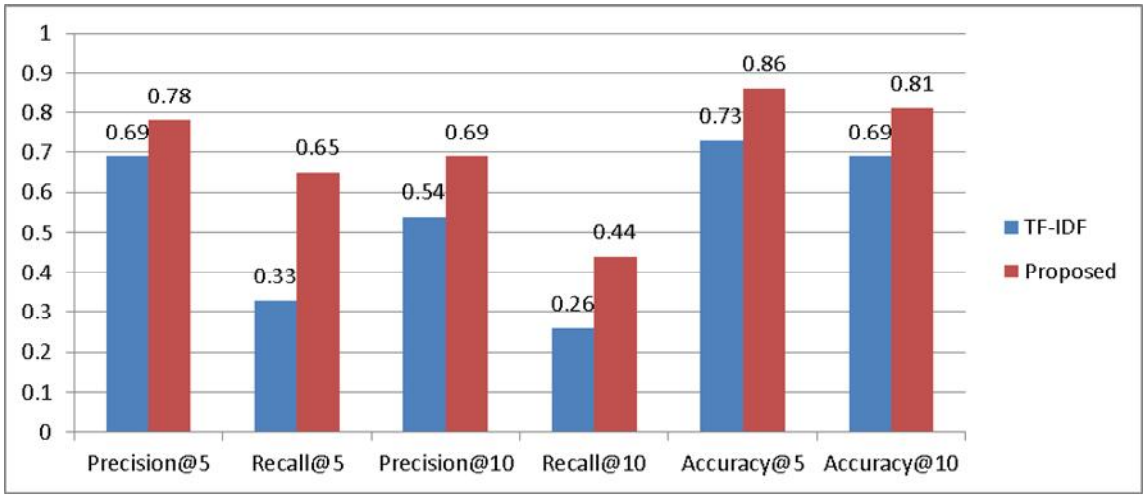

Figure 4.5.0: Precision, Recall and Accuracy at Top 5 and 10 Documents

According to these results provided precision, recall and accuracy were calculated for both methods. The proposed approach to term proximity on TF-IDF has a higher precision compared to the baseline TF-IDF at both top 5 and top 10 documents of 78% and 65% respectively. Recall for the proposed method is slightly lower than the precision. Figure 4.5.0 shows that precision is decreasing as the number of documents retrieved are increased. The proposed method has a higher accuracy score of 86% at 5 documents and 81% at 10 documents which is much better than the baseline method.

4.5.1 Proposed Framework versus Baseline Framework

The proposed framework was tested for precision of docId 235 and DocId 236 shown in table below. User query keywords were ***ploughing*** and ***weeding***. DocId 235 is more relevant and should be ranked higher.

Table 4.5.1: Sample Documents

DocId	**Title**	**Effective land preparation methods**
235	**Body**	Before we engage in planting proper land preparation is very essential. Cattle ploughing is a land preparation method still in use here in Kenya, it not very effective. Tractor driven ploughing would be more effective especially where farm is too large. Most farmers prefer to engage manual laborers in farm weeding rather than use of chemical methods to control it.
236	**Title**	**Event organizing business booming in Kenya**
	Body	Mr. Madjik is the owner of Madjik an event planning company specializing in weedding planning. His company Madjik was started in the year 2012 when Madjik was still in college and has been growing very fast for the last few years. He has been ploughing back his profit into the business, he has used most of the profit ploughed back to market his business..

Figure 4.5.1: Screenshot of Sample Documents Relevance Ranking

The above screenshot in figure 4.5.1 shows that the proposed framework was effective in ranking the two documents effectively compared to the baseline framework. RSV score is used in overall document ranking. DocId 236 was more relevant than DocId 235 which has a smaller RSV score of 0.8835 compared to 0.8858 of DocId 236 as show by table 4.5.2 below.

Table 4.5.2: Sample Documents Relevance Ranking

DocId	cosine	Pair-proximity	Span-proximity	RSV
235	0.88	0.0025	0.001	0.8835
236	0.88	0.0028	0.003	0.8858

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.0 Summary

The main driving force for high acceptance of internet as a source of agricultural information in developing countries is affordability, perceived usefulness, trust and availability of supporting infrastructure. However, as use of internet in support of agriculture is slowly increasing, there is great concern on access of relevant agricultural information especially maize. It was for this reason that the researcher sort to carry out a research based on objectives in chapter one and we conclude by highlighting the achievement of each objective based on our research findings.

5.1 Use of RSS technology in delivering information from websites

The first objective was achieved by exploring the use of RSS technology in delivering information through literature review, where the researcher looked at how RSS works, format of RSS document and the entire process of RSS feed aggregation. Chapter two sections 2.1 to 2.4 gives detailed information.

5.1.1 Factors affecting use of internet-based technology

In order to achieve the second object, we conducted a survey in which technological and economic issues were investigated. Chapter four gives detailed findings from survey. Most of the information obtained from maize farmers greatly helped answer pertinent questions regarding factors that have influenced the use of web-based technology in support of agriculture.

5.1.2 The use of TF-IDF in Retrieval of Relevant Maize Information

To obtain the third objective a detailed study was conducted, through related work the researcher was able to identify different ways TF-IDF has been used in information retrieval. Weaknesses and strengths of Term Frequency Inverse Document Frequency were identified. There was a need to improve on the relevance of the top-k documents returned by TF-IDF and for that a new approach to term proximity was implemented. The use of TF-IDF and term proximity on the framework has been clearly described in section 2.14 in Chapter 2.

5.1.3 Design of a framework for aggregating and retrieving relevant information

By studying related work on use of TF-IDF and Term Proximity by different researchers it was possible to identify existing gaps and strengths on the use of TF-IDF and Term Proximity. The researcher drafted a framework that is clearly described in Chapter 2 section 2.14. The purpose of going out to collect data was to identify factors that affect use of internet-based technology and information needs of maize farmer. Chapter 4 provides information obtained from the survey.

It is evident from the research findings that there is a serious need to provide a framework that will be able to aggregate and retrieve relevant maize information.

5.1.4 The implementation and test of the framework

The framework was integrated into a web application and to ensure effective implementation there was need to investigate technology preference and usage. Farming information was also sought since there was need to understand farmers' agricultural

information preferences in order to ensure implementation was oriented towards maize farming. The results are presented in Chapter 4 section 4.3.0 and section 4.4.0.

For the purpose of investigating if the proposed approach of integrating TF-IDF with our term proximity approach was effective compared to the baseline TF-IDF an experiment was conducted and the results are presented in chapter 4 sections 4.5. User queries were created and given to the user groups where all documents returned after query executions were inspected for relevance. The relevant and irrelevant documents returned after query executions were recorded and precision was calculated for TF-IDF as the baseline and the proposed method, the results show improved precision.

5.2 Recommendations

The findings described in chapter four it was evident that enough factors exist that encourage use of internet-based technology by maize farmers. Internet based technologies have been accepted as trusted source of agricultural information. Majority of maize farmers have not been able to easily access relevant or even adequate information, there is a need for a framework that can be able to aggregate and retrieve most relevant information from different websites based on users query preferences. There are a number of other challenges that have been experienced by maize farmers while accessing relevant information most of the challenges can easily be addressed effectively.

An experiment was done to test the performance of the proposed method and the baseline TF-IDF, from the results that were obtained it is clear that the approach proposed on this thesis has improved precision on top-k documents returned by TF-IDF

hence validating the framework. The framework is a very resourceful tool for retrieval of relevant information and it is recommended to agricultural organization providing information to farmers as it is able to achieve high precision. We also recommend it for use in any Information System meant for retrieving most relevant information to user based on user preferences.

5.3 Further Research

The results of this research have shown that the approach discussed in this thesis is able to perform much better than the baseline TF-IDF. The results have shown evidence of the potential power of our approach to term proximity and it is quite clear that a proximity scoring function should be included in any information retrieval model to give a significant contribution to improvement of relevance in top-*k* documents.

In the future studies the framework can be improved to consider query term synonyms in overall relevance ranking feature and information retrieval.

The sample used in this study consisted of a limited number of farmers due to budgetary and time constraints. This may have been introduced some bias in our research findings. We therefore recommend for more thorough nationwide research in order to explore further the factors that affect use of internet-based technologies in agriculture for improved integration with our framework.

REFERENCES

1 Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirement for a Co-citation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. Journal of the American Society for Information Science and Technology, 54(6), 550-560.

2 Ajani, E. N. (2014). Promoting the Use of Information and Communication Technologies (ICTs) for Agricultural Transformation in Sub-Saharan Africa: Implications for Policy. Journal of Agricultural & Food Information, 15(1), 42-53.

3 Asman, P., Cannon, S., Sommo, C. (2010) Extending RSS to Meet Central Bank Needs. In: Proceedings of the International Conference on Dublin Core and Metadata Applications. Dublin Core Metadata Initiative, Pittsburgh.

4 Babbie, E. R. (1990). Survey research methods (2nd ed.). Belmont, CA: Wadsworth.

5 Banati, H.,Bedi, P., & Grover, P.S(2006). Evaluating Web Usability from the User's Perspective. *Journal of Computer Science 2 (4): 314-317, 2006.* ISSN 1549-3636.

6 Brewington, B & Cybenko, G. (2000) "How Dynamic is the Web?", Proceedings of the 9th International World Wide Web Conference, pages 257–276.

7 Broschart, A..(2012). Efficient Query Processing and index tuning using proximity scores.Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR.

8 Brugger, F . (2011).Mobile Applications in Agriculture. Syngenta Foundation.

9 Buttcher, S. Clarke,C and Lushman, B.(2006). Term proximity scoring for ad-hoc retrieval on very large text collections. In SIGIR '03: Proceedings of the 26nd annual

international ACM SIGIR conference on Research and development in information retrieval.

10 Chisenga, J. (2006). Development and use of Institutional repositories and open Access Archives for Research and National Development in Africa: Opportunities and Challenges. FAO-Regional Office for Accra, Ghana.

11 Cummins, R and O'Riordan, C.(2009). An axiomatic study of learned term-weighting schemes. Learning in a Pairwise Term-Term Proximity Framework for Information Retrieval - SIGIR'09, Boston, Massachusetts, USA

12 Dan, S., Alan, M., Ansley, P., Peter, D. (2004), *FeedTree: Sharing Web Micronews with Peer-to-peer Event Notification*, Department of Computer Science, Rice University, Houston, TX.

13 Dan, A.G. & Shani, G.(2009)A survey of accuracy evaluation metrics of recommendation tasks. Journal of Machine Learning Research.

14 De Silva, Harsha & Dimuthu Ratnadiwakara (2008), 'Using ICT to reduce transaction costs in agriculture through better communication: A case-study from Sri Lanka', mimeo, 20.

15 Fannin, B. L. & Chenault, E. A. (2005). Blogging agricultural news: A new technology to distribute news real-time. Paper presented at the Southern Association of Agricultural Scientists Conference. Little Rock, AR

16 FAO (2011).Situation Analysis: Improving Food Safety In The Maize Value Chain In Kenya

17 Freeman, R.(2009). “Web Feed Clustering and Tagging Aggregator Using Topological Tree-Based Self-Organizing Maps” in Proceedings of the Tenth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL’09) Springer, pp. 368-375

18 Gay, L.R.(1981). Education Research: Competencies for Analysis and Application. Charcles E. Mairill Publishing Company A. Bell & Howell Company. Collumbus, Toronto, London.

19 Gebre, B. G., Zampieri, M., Wittenburg, P., & Heskes, T. (2013). Improving Native Language Identification with TF-IDF weighting. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 216-223).

20 Guo, L. and Peng, Q.K. (2013).A Combinative Similarity Computing Measure for Collaborative Filtering-Applied Mechanics and Materials, Volumes 347-350,pg 2919

21 Hawking, D,A and Thistlewaite, P, B.(1995). "Search For Meaning With The Help of A Padre". Proceeding of the Third Text Retrieval Conference, pages 257-268. US Department of Commerce, NIST Special Publication 500-225.

22 Hendron, J. (2008). RSS for Educator: Blogs, Newsfeeds, Podcasts and Wikis in classroom. International Society for Technology in Education (ISTE). USA, Washington DC. ISBN 978-1-56484-239-8.

23 Indrawan, M.(1998). “A. framework for Information retrieval based on Bayesian Networks”. Monash University.

24 Isah, H. (2012). Full Data Controlled Web-Based Feed Aggregator .International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 3, June 2012.

25 Jayne, T.S., T. Yamano, J. Nyoro, and T. Awour. (2001). *Do Farmers Really Benefit From High Food Prices Balance Rural Interests in Kenya's Maize Pricing and Marketing Policy*. Working Paper 2b, Tegemeo Institute, Nairobi, Kenya.

26 Jones, S. K(2004) "IDF term weighting and IR research lessons," Journal of Documentation, 60(6), pp. 521-523

27 Kantor, A (2007), "Real Simple Syndicate needs to add some complexity",Available: http://www.cyberspeak.com/Real Simple Syndicate needs to add some complexity/ (Accessed: 2012, August 5)

28 Kathuri, N.J. and Pals, A.D. (1993). Introduction to Education Research. Education Media Centre, Egerton University

29 Kedrosky, P. (2004, June). Feeding time. Harvard Business Review, 18-19

30 Kim, S.J. & Lee, S.H. (2005). "An Empirical Study on the Change of Web Pages". Proceedings of the Seventh Asia-Pacific Web Conference, pages 632–642.

31 Kizito, A.M. (2011). The Structure, conduct and Performance of Agricultural Market Information Systems in Sub-Saharan Africa. Michigan State University.

32 Krovetz, R.(1993). "Viewing morphology as an inference process". Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.,191-202.

33 Liu, M and Yang, J. (2012). An improvement of TF-IDF weighting in text categorization. International Conference on Computer Technology and Science. DOI 10.7763/IPCSIT V47.

34 Luhn, H. P.(1958), 'The automatic creation of literature abstracts', IBM Journal of research and development, vol.2, no. 2, pp. 159-165

35 Nagao, D. (2008). Web Content Recommender System on RSS using weighted TFIDF University of Aizu, Graduation Thesis.

36 Mukai, M and Aono, M.(2005). "A Prototype of Content-based Recommendation System based on RSS," Tech. Rep. 2005-FI-80, IPSJ SIG.

37 Maron, M. E., Kuhns ,J. L., and Ray ,L. C.(1959). 'Probabilistic indexing: A statistical technique for document identification and retrieval', Thompson Ramo Wooldridge Inc, Los Angeles, California, Data Systems Project Office, Technical Memorandum 3,.

38 Melucci, M and Orio, N.(2003). "A novel method for stemmer generation based on hidden Markov models". Proceedings of the twelfth international conference on Information and knowledge management, 131-138

39 Monz, C.(2004). Minimal span weighting retrieval for question answering.In Rob Gaizauskas, Mark Greenwood, and Mark Hepple, editors, Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering, pages 23–30, 2004

40 Muriithi,B., Gikandi,A., Ogalleh S.(2008).Information Technology for Agriculture and Rural Development in Africa: Experiences from Kenya.

41 Ndwiga, J., J. Pittchar, P. Musyoka, D. Nyagol, G. Marechera, G.Omanya, and M. Oluoch. (2013). Integrated Striga Management in Africa Poject. Constraints and opportunities of maize production in Western Kenya: a baseline assessment of striga extent, severity, and control technologies. Integrated Striga Management in Africa (ISMA). 34 pp.

42 Paice, D. (1994). "An evaluation method for stemming algorithms". Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. 1994, 4250.

43 Ramos, J. (2003) Using TF-IDF to Determine Word Relevance in Document Queries Rutgers University.

44 Rasolofo, Y and Savoy, J.(2003). Term proximity scoring for keyword-based retrieval systems. In Proceedings of the 25th European Conference on IR Research (ECIR 2003), pages 207–218.

45 Salton, G. and Buckley, C. (1988). Term-weighing approaches in automatic text retrieval. In Information Processing & Management, 24(5): 513-523.

46 Sanderson, M and Croft, W. B. (2006). The History of Information Retrieval Research.

47 Saravanan, R. (2010). ICTs for agricultural extension. In: R. Saravanan (Ed.), *Global experiments, innovations and experiences*. New Delhi, India: New India Publishing Agency.

48 Song, R. Taylor, M. Wen, J. Hon, H. and Yu, Y.(2008). Viewing term proximity from a different perspective. vol 4956, pp. 346357, Springer Berlin /Heidelberg, 2008.

49 Soucy,P and Mineau, G,W.(2008).Beyond TFIDF Weighting for Text Categorization in Vector Space Model.

50 Tao,T. and Zhai,C.(2007). An exploration of proximity measures in information retrieval. In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 295–302, New York, NY, USA,. ACM.

51 Tseng, C., Ng, P, (2007). "Precisiated information retrieval for RSS feeds", Information Management & Computer Security, Vol. 15 Iss: 3, pp.184 – 200.ISSN: 0968-5227.

52 Wartena, C., Brussee, R. and Slakhorst,W.(2010).Keyword Extraction using Word Co-occurrence.

53 Yates,R. A and Neto,B.(1999). A. Modern Information Retrieval. ACM Press, New York, NY, USA.

54 Yin, R. K. (2003). Case study research: Design and methods (3rd ed.). Thousand Oaks, CA: Sage

55 Zeki, C. (2004), *What is RSS and How can it Serve Libraries?*, Istanbul Technical University, Istanbul.

56 Zhang, Q., Zhang, L., Dong S. and Tan J. (2005) Document indexing in text categorization Proceedings of 2005 International Conference on Volume 6, Issue , 18-21.

57 Zhao, J and Yun, Y.(2009). A proximity language model for information retrieval. In SIGIR ’09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in informationretrieval,pages291-298, NewYork, NY, USA. ACM.

58 Zjip, W. (1994) Improving the Transfer and Use of Agricultural Information: A Guide to Information Technology. World Bank Discussion Paper 247, Washington: World Bank.

59 Zhu, M., Shi, S., MingJing, L and Wen, J. (2007). Effective Top-K Computation in Retrieving Structured Documents with Term-Proximity Support. ACM 978-1-59593-803-9/07/0011.pages 771-780.

APPENDICES

QUESTIONNAIRE INSTRUMENT

FOR

FRAMEWORK FOR AGGREGATING AND RETRIEVING RELEVANT INFORMATION IN SUPPORT OF MAIZE PRODUCTION

SERIAL

No…..…

	DD	MM	YY
Date of interview	☐☐	☐☐	☐☐

This research is purely academic; information collected through this research will be confidential and will solely be used for that purpose.

I wish to communicate information about the survey to you. Should you be interested, please indicate your email address on the first page of the questionnaire.

Please take a moment of your time to answer the survey questions. I will appreciate your frank and critical response to this questionnaire.

INSTRUCTIONS

1. This Questionnaire consists of **3 Sections**. Please answer all questions in each section
2. Do not indicate you **Name** on the questionnaire

3. Make sure that you tick within the box.

SECTION A: BACKGROUND INFORMATION

1) Select your Gender

☐ Male ☐ Female

2) Age bracket

☐ 18yrs - 30yrs ☐ 31yrs-40yrs ☐ 41yrs-50yrs ☐ 51yrs-60yrs ☐ Above 60yrs

3) Highest education level

☐ No Schooling Completed ☐ Certificate ☐ Diploma

☐ Bachelors ☐ Masters and Above

4) Marital Status

☐ Married ☐ Divorced ☐ Widowed ☐ Single ☐ Never Married

5) How would you describe your computer skills?

☐ Poor ☐ Moderate ☐ Good ☐ Excellent

6) How would you describe your internet skills?

☐ Poor ☐ Moderate ☐ Good ☐ Excellent

SECTION B: FACTORS INFLUENCING USE OF INTERNET TECHNOLOGY.

7) How would you describe your monthly income?

☐ Less than 26,000 ☐ 26,000 to 50,000 ☐ 51000 to 75000

☐ 76000 to 10000 ☐ over 100000

8) Do you have formal training on the use of computers?

☐ YES ☐ NO

9) Do you have formal training on the use of internet?

☐ YES ☐ NO

10) Trust: I trust agricultural information delivered through Internet technology.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

11) Perceived usefulness: Internet based technology has been useful in providing necessary agricultural information

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

12) I have easy access to the necessary internet infrastructure for access to agricultural information

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

13) Do you pay for the internet you use to access agricultural information?

☐ YES ☐ NO

14) If you indicated YES in question 13 how do you rate the cost of your internet connection?

☐ Not expensive ☐ Fairly Expensive ☐ Very expensive

SECTION C: FARMING INFORMATION

15) How long have you been a maize farmer?

☐ Less than 5yrs ☐ 5yrs-10yrs ☐ 11yrs-15yrs ☐ 16yrs-20yrs ☐ 21yrs>

16) What size of farm do you grow maize crop?

☐ Less than 5 acres ☐ 5acres-10acres ☐ 11acres-15acres ☐ 16acres-20 ☐ 21acres>

17) Type of Farming. What type of maize farming do you do?

☐ Subsistence ☐ Commercial ☐ All Above

If Commercial or Both where do you sell your maize harvest?

_____

_____

18) If you answered Commercial or Both do you experience problems transporting produce to market place? YES ☐ NO ☐

If you answered YES what problem do you experience?

_____

_____

19) What type of land preparation do you mostly use

☐ Animals ☐ Tractors ☐ manual labour ☐ all above

20) What type of maize seeds do you grow on your farm?

☐ Local ☐ Hybrid

If you answered Hybrid indicate the maize variety you plant on your farm.

_____

_____

21) How often do you use fertilizer in your maize farming?

☐ Very often ☐ Often ☐ Rarely ☐ Not at all

If you did not indicate Not at all in question 21, what kind of fertilizer do you use?

_____

_____

22) How often do you use pesticide in your maize farming?

☐ Very often ☐ Often ☐ Rarely ☐ Not at all

If you did not indicate Not at all in question 22, what pesticides do you use?

_____

_____

23) Do you use herbicides in weed control?

☐ YES ☐ NO

If YES what herbicide do you use?

_____

_____

24) Indicate any pests or diseases that have affected your maize farming?

_____

_____

25) How often do you use weather forecasting information before planting your maize?

☐ Very often ☐ often ☐ Rarely ☐ Not at all

Indicate your source of weather information_____

26) What was your total maize yield in the last season?

☐ <10 bags ☐ 11-20 bag ☐ 21-30 bags ☐ 31-40 bags ☐ > 40 bags

27) How do you intent to achieve more maize yield in the next planting season?

_____

_____

28) Do you experience any maize storage problems?

☐ YES ☐ NO

If YES what king of problems do you experience?

_____

_____

29) Do you have other sources of agricultural information besides internet?

☐ YES ☐ NO

If YES indicate your other source of information_____

30) If you indicated YES in 29.I find internet information timelier than my other sources of information.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

31) What kind of maize information are you mostly interested in?

_____

_____

SECTION D: TECHNOLOGY USAGE

32) What device do you prefer to use while accessing maize information online?

☐ Desktop ☐ Laptop ☐ Mobile phone ☐ I-pad/ Tablet

33) Among these electronic devices used to access internet, select devices you own.

☐ Desktop ☐ Laptop ☐ Mobile phone ☐ I-pad/ Tablet

34) Through which of the means listed below do you currently receiving maize information?

☐ Through SMS ☐ Through email ☐ website

35) Which would be the your most preferred mean of receiving maize information?

☐ Through SMS ☐ Through email ☐ website

36) How many times within a month do you access agricultural information online?

☐ Less than 5 ☐ 6 to 10 ☐ 11-20 ☐ 21-30 ☐ Over 30

37) The websites i use to access agricultural information are user-friendly

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

38) I get enough maize information from agricultural websites i access.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

39) Information I access on agricultural websites/blogs is completely relevant to my information needs.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree