The Feed Analyzer: Implementation and Evaluation

Sahar Bazargani, Julian Brinkley, Nasseh Tabrizi
Department of Computer Science
East Carolina University, Greenville NC 27858
USA
Bazarganis10@students.ecu.edu, Brinkleyju11@students.ecu.edu, Tabrizim@ecu.edu

Journal of Digital Information Management

ABSTRACT: *Modern society is producing more information at a faster pace than ever before; data which is increasingly used to solve a variety of real world problems. But given the vast size of the data involved and the resource intensive nature of rapid large-data processing, the need for more advanced methodologies in this regard is growing. This phenomenon has given rise to the term 'Big Data' which references the types of data intensive problems that are typically beyond the ability of traditional tools and methodologies to effectively manage. This paper outlines the use of agent-oriented software engineering methodologies and commercial cloud technology in the development of a system designed to address one of these many Big Data problems. The Feed Analyzer is a conceptual search based Web feed aggregation system deployed to Microsoft's Windows Azure cloud platform. We discuss the implementation and evaluation of this system with the primary goal of contributing to the body of knowledge a practical means by which a Big Data problem may be addressed using cloud technologies.*

Subject Categories and Descriptors
H.2 [Database Management]; H.3.5 [**Online Information Services**]: Web-based services

General Terms
Bigdata, Cloud Computing, Web Aggregation System

Keywords: Agent Based Systems, Design Patterns, Cloud Computing, Big Data

Received: 12 August 2014, Revised 30 September 2014, Accepted 5 October 2014

1. Introduction

The evolution of data management technology has aided the ability to collect, store, access and analyze data about the most minute to the most significant of activities. But the world's increasing computerization and interconnected nature means that more of this information is being produced faster than ever before and used in a multitude of contexts. Identity and behavioral information for instance, aggregated from any number of disparate sources, is increasingly used to aid law enforcement in identity deception scenarios [1] while also providing predictive models of potential criminal activity [2]. This information may be similarly used to aid in fraud prevention both for online as well as traditional "*brick and mortar*" retailers. This overall phenomenon has given rise to the term '*Big Data*' which generally refers to often multi-sourced volumes of aggregated information that are "*too big, too fast, or too hard for existing tools to process*" [3].

The rise of these Big Data problems has resulted in the need for tools and methodologies that address the challenges inherent in the rapid processing of these large volumes of aggregated information. This paper, the second in a series, documents an attempt to address one of these Big Data problems within the context of a Web feed data aggregation system call the Feed Analyzer. The Feed Analyzer implements a type of semantic or concept-based search capability with the aim of more broadly identifying potential feeds of user interest while also attempting to minimize the return of extraneous data with minimal relevance to the user's identified topic(s) of interest. The development of this application using agent-based software engineering methodologies and cloud technologies provides evidence to support the contention that both may aide in the solution of Big Data problems of this type.

2. Background: Website Data Feeds

A Website data feed or news feed is a document which summarizes the selected content of a website and contains references to each discrete content item's underlying location; typically via a Web link or hyperlink [4]. As these often XML-based [5] feeds are designed to be machine readable they are typically consumed by stand-alone or Web browser based feed readers or feed aggregators [6] via a dissemination process referred to as syndication. From a functionality perspective these feed aggregation systems characteristically enable users to:

- Subscribe to one or more website news feeds of interest.
- View the collected content items from each feed displayed in a human-friendly format.
- Access source content for each feed item via an included Web link.

The time saving benefits of feed aggregation are readily apparent when one considers the volatile nature of online content and the time intensive nature of manual data perusal.

News websites and Web logs feature content that is highly volatile by definition; newsworthy events may occur at any time and Web logs are updated at the whim of their publishers. As a result, manually checking a single source for new or updated content is often a significantly time consuming task, a task exacerbated by the average user's desire to receive information from more than one source. The ability to automate this data collection process potentially enables a user to access more information of interest, in a shorter time span, while ensuring this information is more current and up-to-date than would be otherwise likely.

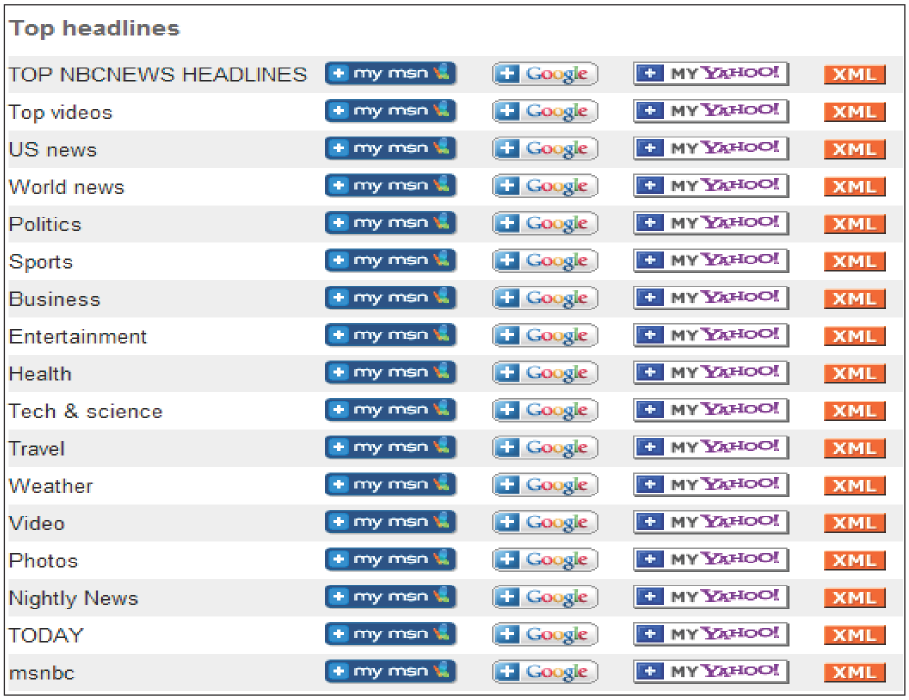
While a number of feed formats exist, the XML-based Really Simple Syndication (RSS) [7] and Atom [8] formats are the most widely used, with RSS being the more popular. Both formats contain hierarchically structured information though significant differences exist regarding the structure of each document. While an in depth comparison of the these two formats is outside the scope of this document, the most significant distinction between the RSS and Atom formats is arguably the feed content model. The Atom format requires the explicit identification of the content type, HTML or XHMTL for instance, whereas the RSS format imposes no such restriction and lacks a descriptive mechanism [8].

2.1 Problems with Traditional Feed Aggregators

But despite the aforementioned time savings that feed aggregators make possible they still generally possess a significant drawback; information overload. While traditional Web data feed aggregators are undoubtedly effective in collecting feed data they are generally deficient in terms of their search and filtering mechanisms. While this may appear to be a relatively trivial deficiency, this limitation in extreme circumstances undermines the time saving purpose of data aggregation in that it may quickly lead to information overload. While checking individual websites for updated content is generally a time consuming task, it is readily apparent that manually mining a large volume of aggregated feed data for the potentially limited information of interest can be equally time consuming at a minimum. To mitigate this potential for information overload, feeds are often categorized at the source into popular categories, as is the case on the news oriented website NBCNews.com [9] and illustrated in Figure 2.1. But a brief review of this categorized feed data using the Google Reader [10] and two NBCNews.com RSS feeds reveals the deficiency of this method and current feed applications generally.

As of the time of this writing the NBC News Sports feed added on average 258.3 new posts per week and the Technology feed an average of 158 new posts per week. But what if a user is only interested in sports stories that directly pertain to American professional soccer? Or perhaps technology items that focus on the newest smartphone products? In a best case scenario using the Google Reader product and most current market offerings, a user would need to perform a keyword-based search on the more than 400 combined weekly items contained within these two feeds in an attempt to manually identify the specific information of interest (Figure 2.2). Unfortunately, these keyword-based searches are often deficient in terms of identifying all potentially relevant feed items. As illustrated in Figure 2.2 a search using the phrase "*electronic commerce*" for instance, against the NBCNews.com Business Feed, returns zero (0) news items whereas a search using "*e Business*" returns nine (9) as show in Figure 2.3. Therefore this decidedly time consuming and inefficient process, which undermines the central purpose of feed aggregation systems, may still produce results that omit desired information.

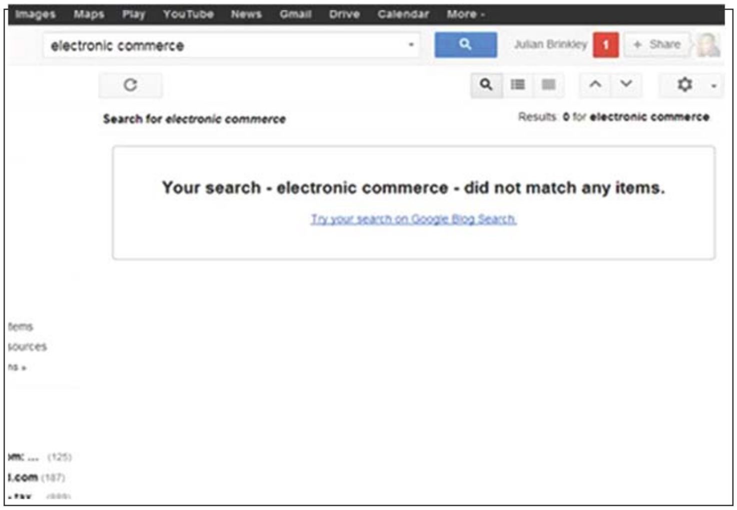
Concept-based or semantic search as implemented within the Feed Analyzer system has been designed to address both issues by employing a method of filtering and categorization that is significantly more effective in terms of identifying information of user interest. The conceptual search capability, as implemented, has been designed to more broadly identify information of interest while also returning more refined results. Implementing this capability however is complicated by the large volume of data that must be searched using this method. As a result concept-based Web feed aggregation can be viewed as a classic Big Data problem:



Top headlines

TOP NBCNEWS HEADLINES	+ my msn	+ Google	+ MY YAHOO!	XML
Top videos	+ my msn	+ Google	+ MY YAHOO!	XML
US news	+ my msn	+ Google	+ MY YAHOO!	XML
World news	+ my msn	+ Google	+ MY YAHOO!	XML
Politics	+ my msn	+ Google	+ MY YAHOO!	XML
Sports	+ my msn	+ Google	+ MY YAHOO!	XML
Business	+ my msn	+ Google	+ MY YAHOO!	XML
Entertainment	+ my msn	+ Google	+ MY YAHOO!	XML
Health	+ my msn	+ Google	+ MY YAHOO!	XML
Tech & science	+ my msn	+ Google	+ MY YAHOO!	XML
Travel	+ my msn	+ Google	+ MY YAHOO!	XML
Weather	+ my msn	+ Google	+ MY YAHOO!	XML
Video	+ my msn	+ Google	+ MY YAHOO!	XML
Photos	+ my msn	+ Google	+ MY YAHOO!	XML
Nightly News	+ my msn	+ Google	+ MY YAHOO!	XML
TODAY	+ my msn	+ Google	+ MY YAHOO!	XML
msnbc	+ my msn	+ Google	+ MY YAHOO!	XML

Figure 2.1: Selected feed categories on the news website NBCnews.com



electronic commerce

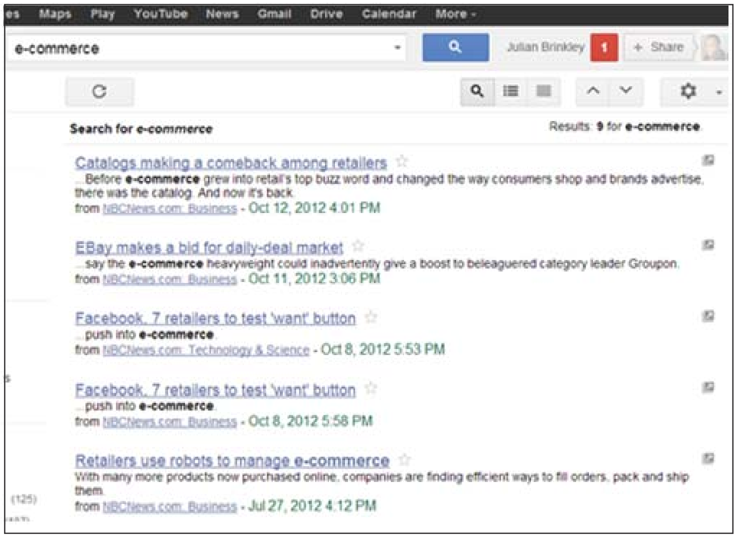
Search for *electronic commerce*

Results: **0** for **electronic commerce**

Your search - electronic commerce - did not match any items.

Try your search on Google Blog Search.

Figure 2.2. Keyword based search within the NBCnews.com Business feed for “*electronic commerce*” using Google Reader. The search term returns zero (0) items



e-commerce

Search for *e-commerce*

Results: **9** for **e-commerce**

Catalogs making a comeback among retailers
...Before **e-commerce** grew into retail's top buzz word and changed the way consumers shop and brands advertise, there was the catalog. And now it's back.
from NBCNews.com: Business - Oct 12, 2012 4:01 PM

EBay makes a bid for daily-deal market
...say the **e-commerce** heavyweight could inadvertently give a boost to beleaguered category leader Groupon.
from NBCNews.com: Business - Oct 11, 2012 3:06 PM

Facebook, 7 retailers to test 'want' button
...push into **e-commerce**.
from NBCNews.com: Technology & Science - Oct 8, 2012 5:53 PM

Facebook, 7 retailers to test 'want' button
...push into **e-commerce**.
from NBCNews.com: Business - Oct 8, 2012 5:58 PM

Retailers use robots to manage e-commerce
With many more products now purchased online, companies are finding efficient ways to fill orders, pack and ship them.
from NBCNews.com: Business - Jul 27, 2012 4:12 PM

Figure 2.3: Keyword based search within the NBCnews.com Business feed for “*e-commerce*” using Google news. The search term returns nine (9) items

- **Too big**: Current estimates indicate that the Web is composed of over half a million websites and nearly a trillion pages of content; numbers that are expected to continually increase [11], [12]. The growth of Web feed data will logically increase as the amount of online information grows.
- **Too fast:** Given that the central purpose of feed aggregation systems is to eliminate the need to manually check websites for updated content, results must be returned quickly so as to avoid negating the time saving benefits of feed aggregation.
- **Too hard**: The large volume of Web feed data coupled with the need for rapid processing results in a complex problem that is difficult for traditional tools and methodologies to address.

3. System Implementation

From a high level perspective the Feed Analyzer application has been conceived as a complex system capable of:

- Retrieving data feed information from a large number of sources (websites).
- Converting this information to the system internal format.
- Indexing the converted data in order to facilitate searching, filtering, and analysis.
- Extending a Web service to facilitate user interaction.

This functionality, in and of itself, is neither new nor novel as it shares commonalities with the majority of website feed aggregators currently on the market. These existing products provide a group of features that facilitates information access and conceivably improves the users' experiences in accessing this information. The Feed Analyzer is differentiated from current market offerings however by its implementation of conceptual search capability aided by an agent-based design and the utilization of cloud technology. These three factors have enabled the Feed Analyzer to produce results that are decidedly more effective than keyword based feed aggregators; identifying feeds of user interest that would otherwise be overlooked as documented in section 4.

3.1 Conceptual Search Methodology and Implementation

The majority of feed aggregation systems on the market, to include browser based systems, categorize feed information by either publication date or feed source while allowing this information to be filtered by keyword or subject. As the previous example demonstrated however, these limited features do not necessarily satisfy the needs of users who often require the actual semantic or "*meaning-based*" search of feed content often referred to as conceptual search capability [13]. The omission or exclusion of this conceptual search capability forces users to sift through feeds manually in an attempt to find stories and other content of interest; a decidedly time intensive task.

3.1.1 The Limitations of Keyword Based Search Capability

Most feed aggregators implement a version of the keyword based search illustrated in Figure 2.3 as an alternative to the manual review of each feed's content. But while this capability is hypothetically sufficient, in practice it produces results that are decidedly lacking given the number of terms that may be used to reference the same underlying information. Consider the subject electronic commerce. According to the Library of Congress Subject Headings (LCSH), an information classification thesaurus that has been actively maintained since 1898, there exist 11 alternate terms or labels for electronic commerce to include eBusiness, eCommerce and internet commerce [14]. As a result, a keyword based search for the term electronic commerce has a significant chance of missing relevant stories and news given that only items that specifically include the term electronic commerce within their content or description will be returned. An attempt to return all potentially relevant information would require a user's awareness of all possible alternatives to the initial search term which might still omit desired results.

And this is by no means an isolated issue confined to only a handful of potential terms. An investigation of the more than 350,000 records within the subject dataset of the LCSH thesaurus has shown that nearly 38% of subjects have at least one semantic alternative; referred to as an Alternate Label (AL) [15]. This means that for more than a third of the subjects within this dataset there exists information that cannot be accessed by searching a single term; some of which possess as many as 30 alternatives.

3.1.2 Conceptual Classification

Conceptual classification refers to a process of information identification and categorization that goes beyond a single term or label. Categorization by concept is essentially an implementation of semantic data classification where the goal is to deduce meaning by determining what information the user was attempting to identify. Conceptual classification asks the question, "*What information is the user attempting to locate*?" In conceptual classification schemes relationships between terms are mapped which enables a system to determine that a user searching for news or stories referencing electronic commerce is also seeking information on eBusiness, eCommerce and other related terms. In short, conceptual classification methodology makes conceptual search capability possible.

3.1.3. Library of Congress Subject Heading Thesaurus

Conceptual classification and the identification of the relationship between terms within the Feed Analyzer system has been based on the subject relationships identified within the Library of Congress Subject Heading thesaurus. While a number of factors contributed to the use of the LCSH thesaurus, its use was logical given that access to this information is free, the thesaurus is updated annually and the relationship between terms identified within the thesaurus has been deemed sufficient for the needs of the Feed Analyzer system.

The LCSH has a multi-parent tree structure of subjects which provides an exhaustive mapping of subjects and their related terms [15]. The relationship between a subject and its alternate terms is illustrated in Fig. 3.1 and outlined below; additional implementation details can be found in our earlier work [16]:

- Main Term (MT): alternately referred to as a Subject or Subject Heading: The term supplied by the user or the user's primary term of interest [17].
- Broader Term (BT): A general term for the MT, representing a super class in the subject hierarchy [17].
- Narrower Term (NT): A more specific term for the MT, representing a subclass in the subject hierarchy [17].
- Related Term (RT): A subject that is semantically related to the MT [17].
- Used For Term (UF) or Alternate Label (AL): An alternative term used to reference the MT [15].

The conceptual design of subject entities as implemented in the Feed Analyzer system is diagrammed in Figure 3.2. All associations in this design are a type of "*Has*" association where it can be said that a Subject Heading (SH) may have multiple Narrower Terms (NT), Broader Terms (BT), Related Terms (RT) and Alternate Labels (AL) [15].

3.1.4 Creation of a Subject Hierarchy in SQL Server

To duplicate the LCHS hierarchy in the Feed Analyzer system it is necessary to both extract this relationship information from the LCHS XML file and store it within the system. A discussion of the former will be eschewed in favor of a more detailed analysis of the more complicated latter process; storage and replication of the subject relationships within a relational database. Recreating the subject and term relationships within the Feed Analyzer system was not an easy task given that relational databases do not support hierarchical structure. SQL Server [18], a commercial relational database management system (RDMS) created by Microsoft, was used to store and represent the subject relationship information.

a sub-tree) involves joining multiple tables and results in generally poor performance. This is especially so with large databases like that used within the Feed Analyzer system which at the time of this writing contains over 350,000 records.

A modified version of the Path Enumeration Model was ultimately chosen due to its relatively superior performance and similarity to the LCSH hierarchical structure. In PEM the full path from the root to each node is stored. As a result, identifying descendants involves searching within the path and not joining tables. Given that defining an index on the path field in the database may improve search performance the performance of the search process in the PEM model is considerably more efficient than in the ALM model or others like it. The path used within PEM is nearly identical to the classification number used within the LCSH.

3.2 Agent-Based System Design and Cloud Deployment

Multi-Agent or agent-based software engineering principles have emerged as a means of solving complex, processing intensive problems like those faced by feed aggregation systems. When coupled with cloud technologies, these agent based systems are increasingly emerging as a potential solution to the difficulties of working with Big Data. Agent based systems break down a single complex problem into multiple problems of reduced complexity that are individually addressed by distinct system agents [20]. These agents in turn communicate the status of active processes which improves the overall function of the system in real time as the distributed workload improves overall processing efficiency [21]. And given that systems designed using multi-agent models readily lend themselves to development as cloud services [22] systems designed using agent-based cloud methodologies benefit from access to on-demand resources as well as potentially enhanced agent communication [23],[24]. To capitalize on these benefits the Feed Analyzer system has been designed using agent-based methodologies and deployed to Microsoft's Windows Azure [24] cloud platform service. Each identified agent was subsequently implemented as the following Window's Azure Cloud Roles as described briefly below:

3.2.1 Manager Worker Role

The Manager is responsible for maintaining fresh feed data within the system by constantly monitoring the age of the data against an update interval defined within the system's app.config file; a value that may be modified independently. Feed sources in need of an update are added to a first-in first-out Azure cloud storage container called the Feed Reader queue. Upon update the age of the feed source data is reset thus beginning the process anew.

3.2.2 Feed Reader Worker Role

Asynchronously, instances of the Feed Reader receive messages from the Feed Reader queue. For each of the feed sources in the queue, this service reads the Web feeds which have been published by that feed source. The Feed Reader reads the XML based RSS document, parses it, and saves the feed channels and feed items in database. Given that the cloud queue allows messages to be read without deleting them, if an instance of the Feed Reader failed for any reason the message will be visible to other instances of the Feed Reader after a defined amount of time (default visibility timeout).

3.2.3 Feed Indexer Worker Role

Processing large datasets is almost universally a time and resource intensive task. Without indexing, this process could be indefinite. Search functionality as implemented within the Feed Analyzer system requires that two large datasets are searched; a subject heading dataset containing more than 350,000 records and a feed item dataset containing more than 80,000 records at the time of this writing. Using any conceivable process searching both datasets to generate reports would be a lengthy process. Implementing an indexing process within the system has improved processing times considerably however.

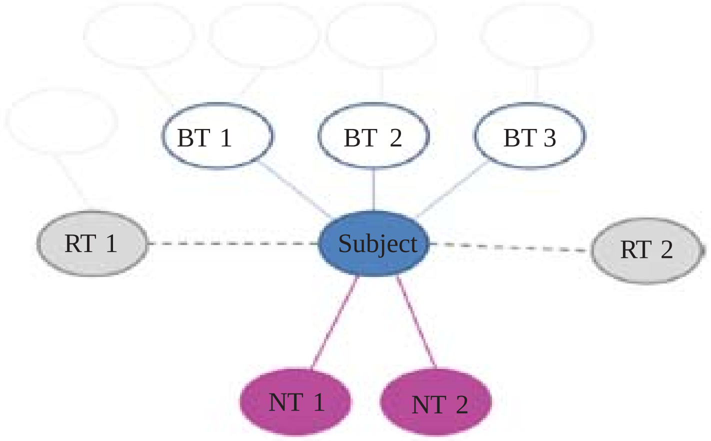


Figure 3.1. Subject relationships within the LCSH thesaurus

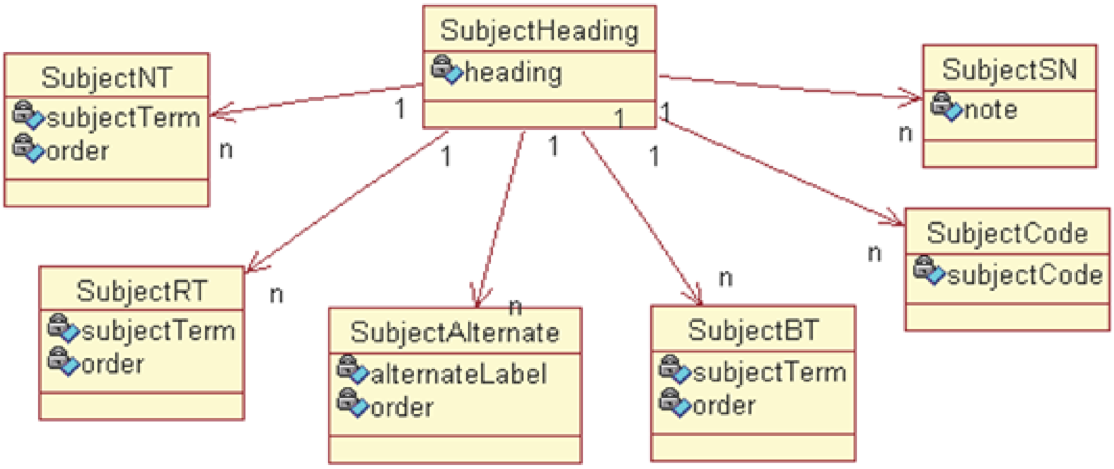
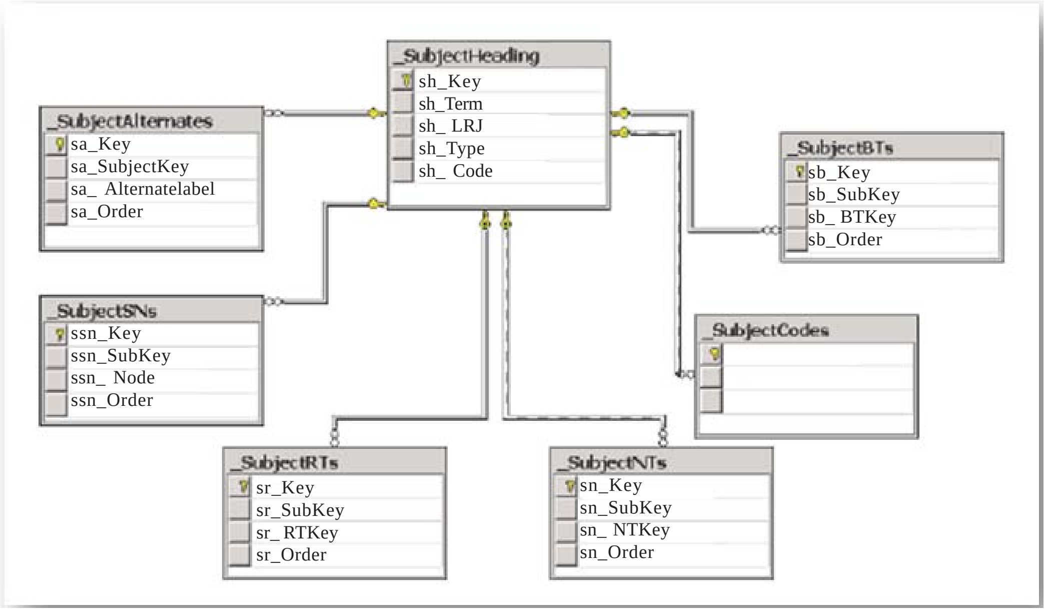


Figure 3.2: Subject relationships of the LCSH thesaurus as implemented within the Feed Analyzer System


_SubjectHeading
sh_Key
sh_Term
sh_ LRJ
sh_Type
sh_ Code

_SubjectAlternates
sa_Key
sa_SubjectKey
sa_ Alternatelabel
sa_Order

_SubjectBTs
sb_Key
sb_SubKey
sb_ BTKey
sb_Order

_SubjectSNs
ssn_Key
ssn_SubKey
ssn_ Node
ssn_Order

_SubjectCodes

_SubjectRTs
sr_Key
sr_SubKey
sr_ RTKey
sr_Order

_SubjectNTs
sn_Key
sn_SubKey
sn_ NTKey
sn_Order

Figure 3.3: Subject Heading Tables and their Relationship after Normalization

3.2.3 Feed Analyzer Service Web Role

This Web role is the sole user interface of the Feed Analyzer system; other services perform back-end processing. Given that one success criteria of the system is an effective user interface, this service will be discussed within the context of the system's effectiveness (section 4).

A more detailed examination of the process used to identify each agent as well as an in-depth examination of each cloud role is available in the first paper of the series which more extensively documents the system design process [26].

4. System Evaluation

As has been previously argued the use of agent-based software engineering methodologies coupled with cloud technologies may aid in solving increasingly common Big Data problems. The effectiveness of the Feed Analyzer system, developed using agent-based methodologies and deployed to the Windows Azure cloud service, provides support for this argument. As the subsequent data will demonstrate, the Feed Analyzer's implementation of conceptual search capabilities has produced results that are significantly more exhaustive and refined than keyword based feed aggregation systems. The system has been able to effectively process two large datasets; a subject dataset which makes possible the system's conceptual search capabilities and a feed dataset that contains regularly updated Web feeds. As of the time of this writing the state of the data in the system was as follows:

- There are almost 100 feed sources in the system.
- The system has downloaded 79,949 feed channels and 80,115 feed items.
- The subject heading table contains 353,307 subject heading records, 235,150 NT (relationship between a subject and its Narrower Term) relations and 14,402 RT (relationship between a subject and its Related Term) relations.
- The index table has 18,245 records.

Given that the ability to search aggregated feed data by concept versus keyword has been identified as the most crucial function of the Feed Analyzer, our results will focus on the effectiveness of the finished system in this regard.

4.1 Usage Scenario

A new user could be expected to interact with the Feed Analyzer system as illustrated in Figures 4.1 through 4.2 and described as follows:

1. Adding Feed Sources: After login and authentication this new user would navigate to "*Feed Sources*". If a particular source of interest has not been added to the system, this user could add this source by providing its Uniform Resource Locator (URL). After a short delay, the Web feeds of this new feed source will be available in the system.

2. Identification of Subjects of Interest: The key differentiating factor between the Feed Analyzer and keyword-search based systems is the ability to identify information of interest by concept as opposed to keyword or phrase. This concept-based search capability is illustrated in Figure 4.1. Users may search the complete LCSH thesaurus to identify subjects of interest and add these subjects to their "*Favorite Subject List*". The system displays the selected subject, any available descriptive information and any alternate terms or labels (AL).

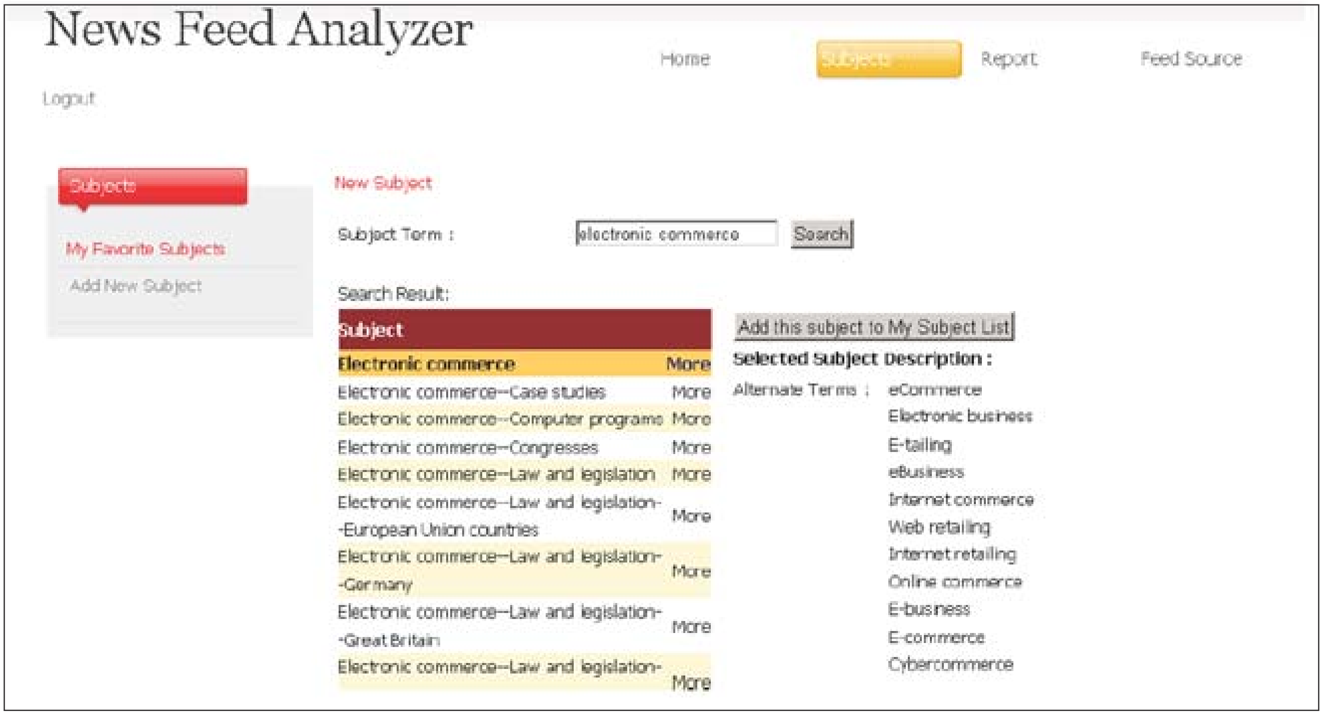
3. Subject Report: The subject report provides a list of the subjects within the user's "*Favorite Subject List*" which have been referenced within the Web feeds themselves as illustrated in Figure 4.2. By selecting a subject from the Subject Term list a user may access additional subject details regarding the Main Term and any Related Terms or Narrower Terms. The report is ranked and sorted by the number of references to the subject(s) within the Web feed dataset. While details regarding the ranking method used are outside the scope of this paper, the method assigns significantly more weight to a user's primary subject selection versus any related or narrower terms.

4.2 Comparison: Conceptual Search vs. Keyword Search

In order to evaluate the effectiveness of conceptual search as implemented in the system a group of 78 subjects was randomly selected from those subjects that possess at least one Alternate Label (AL), Related Term (RT) or Narrower Term (NT). The first 20 of these terms have been selected for graphical illustration in Figures 4.3 through 4.6; feed counts by Subject, RT, NT and AL are provided in Tables I through III.

Our results indicate that subjects with either an AL, RT or NT comprise nearly 50% of the subjects in our thesaurus. For these subjects, implementation of conceptual search capability has produced significantly more effective results in terms of identifying feeds of potential interest. For nearly 90% of the subjects in the test set a conceptual search returns more records than a keyword search as illustrated in Figure 4.3. This is untrue for approximately 10% of the records where the results are identical regardless of the search method used. But these identical results do not necessarily indicate that conceptual search yields no advantage for these subjects. It is entirely possible that as the number of feeds and feed sources in the system increases a performance advantage will emerge.

Alternate labels yield more results than a main subject search, as demonstrated in Table I due to the fact that the main term of a subject in the thesaurus is not always the most common term used to reference that subject. Conceptually, all of the terms that are used to describe a subject have the same value and therefore should return identical result sets.



New Subject

Subject Term : electronic commerce

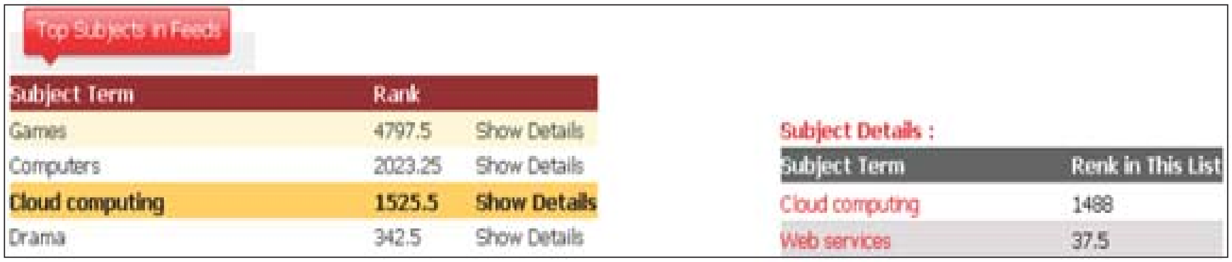
Search Result:

Subject	
Electronic commerce	**More**
Electronic commerce--Case studies	More
Electronic commerce--Computer programs	More
Electronic commerce--Congresses	More
Electronic commerce--Law and legislation	More
Electronic commerce--Law and legislation--European Union countries	More
Electronic commerce--Law and legislation--Germany	More
Electronic commerce--Law and legislation--Great Britain	More
Electronic commerce--Law and legislation-	More

Selected Subject Description :

Alternate Terms : eCommerce
Electronic business
E-tailing
eBusiness
Internet commerce
Web retailing
Internet retailing
Online commerce
E-business
E-commerce
Cybercommerce

Figure 4.1: Feed Analyzer System – Favorite Subject Page



Subject Term	Rank	
Games	4797.5	Show Details
Computers	2023.25	Show Details
Cloud computing	**1525.5**	**Show Details**
Drama	342.5	Show Details

Subject Details :

Subject Term	Renk in This List
Cloud computing	1488
Web services	37.5

Figure 4.2: Feed Analyzer System – Sample Subject Report

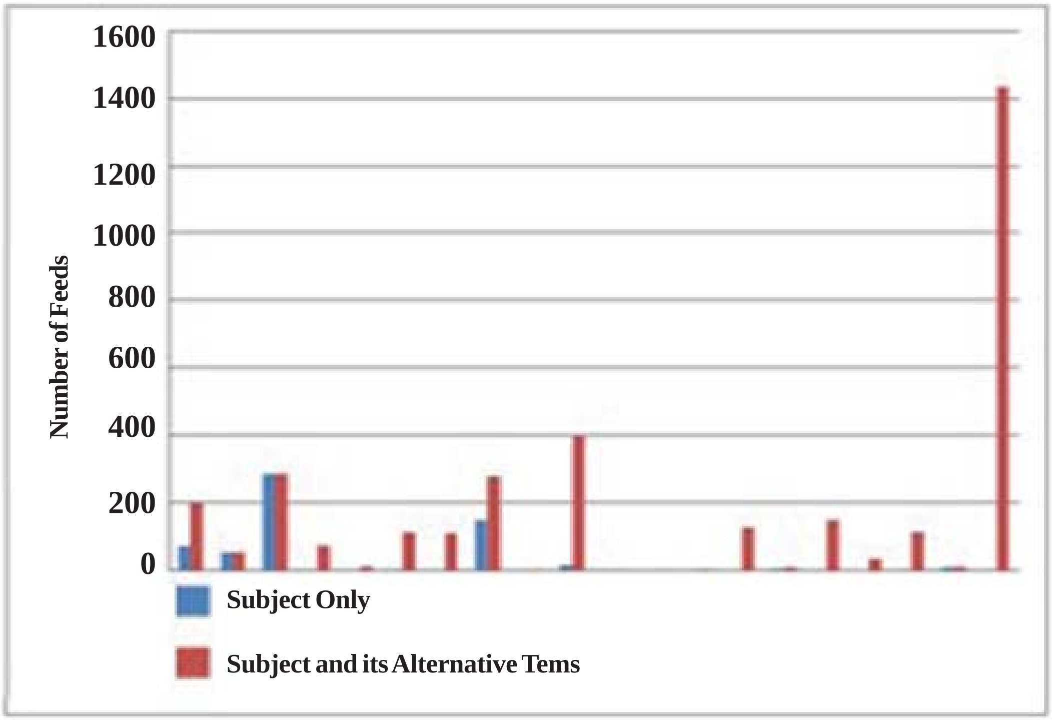


Figure 4.3: Comparison of Search Results (Subject with and without Alternate Labels)

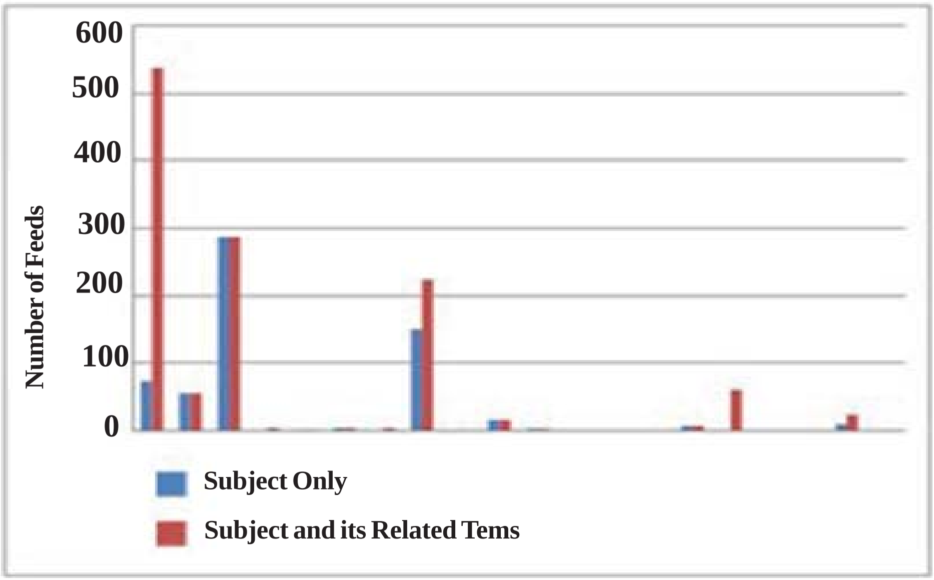


Figure 4.4. Comparison of Search Results (Subject with and without Related Terms)

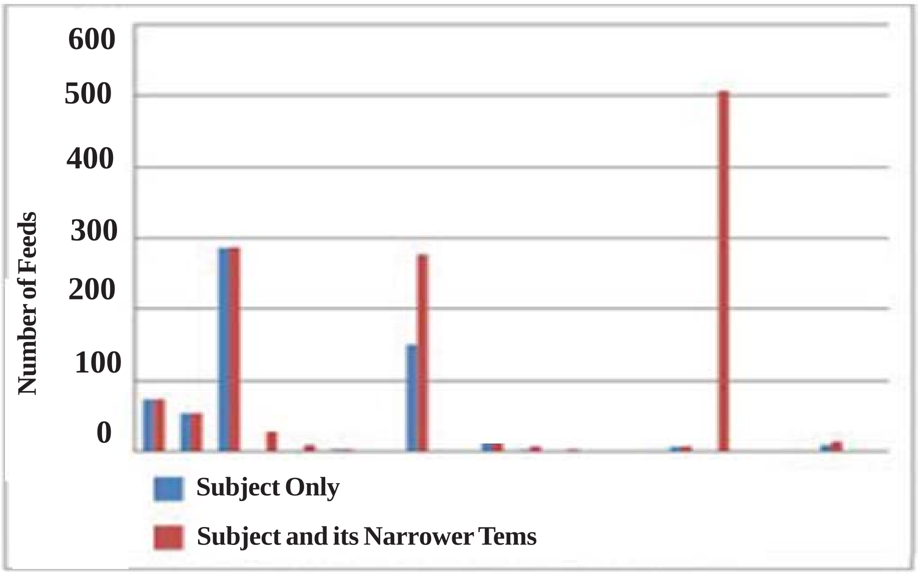


Figure 4.5. Comparison of Search Results (Subject with and without Narrower Terms)

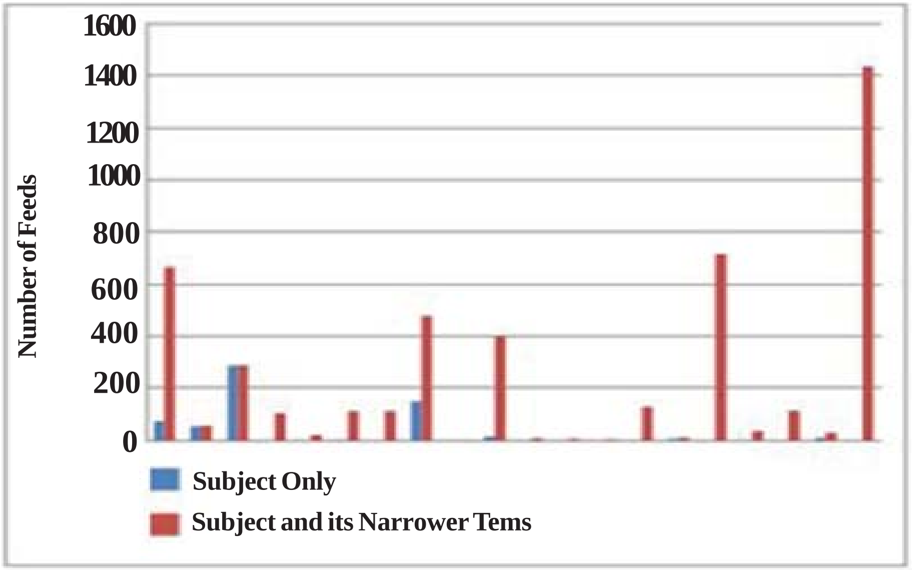


Figure 4.6. Comparison of Subject Search and Conceptual Search

Subjects	Result set size (number of found feeds)		
	Subject Only	Alternate Labels	Subject and Alternate Labels
Architecture	73	128	201
Bankruptcy	54	1	55
Baseball	286	286	286
Belief and doubt	0	74	74
Beverages	1	11	12
Breakfasts	3	109	112
Brunches	0	109	109
Building	150	128	278
Cell interaction	0	2	2
Cell phones	15	389	399
Charities	2	1	3
Child abuse	0	3	3
Chili powder	0	5	5
Church work	0	128	128
Civil rights	6	3	9
Clothing and dress	0	150	150
College sports	0	36	36
Composers	0	114	114
Computer security	9	2	11
Computer software-Development	0	1434	1434

Table 1. Result Set Size, Searching by Subject, Alternate Label (AL) and Combined

Subjects	Result set size (number of found feeds)		
	Subject Only	Alternate Labels	Subject and Alternate Labels
Architecture	73	465	537
Bankruptcy	54	0	54
Baseball	286	0	286
Belief and doubt	0	3	3
Beverages	1	0	1
Breakfasts	3	0	3
Brunches	0	3	3
Building	150	73	223
Cell interaction	0	0	0
Cell phones	15	0	15
Charities	2	0	2
Child abuse	0	0	0
Chili powder	0	0	0
Church work	0	0	0
Civil rights	6	0	6
Clothing and dress	0	60	60
College sports	0	0	0
Composers	0	0	0
Computer security	9	14	23
Computer software-Development	0	0	0

Table 2. Result Set Size, Searching by Subject, Related Term (RT) and Combined

5. Conclusion and Future Work

This paper demonstrates that multi-agent methodologies, when combined with cloud technology, can be used to effectively solve "*Big Data*" problems. Our specific conclusions regarding this premise as pertains to the Feed Analyzer system are as follows:

- Using Semantic search capability as implemented within the Feed Analyzer system is significantly more effective at identifying information of interest when compared with keyword based searches. This addresses one of the central weaknesses of current feed aggregators – information overload caused by deficient filtering mechanisms.

- This semantic search capability has been made possible largely by the use of cloud technology given the need of the system to process two large datasets; a subject dataset which maintains subject term and relationship data and a feed dataset that contains regularly updated Web feed content.

- Development of the Feed Analyzer as a cloud service has increased scalability while also aiding in the forward compatibility of the system. While the system currently only consumes feeds in RSS format, modification for compatibility with Atom or any future format would require minimal disruption to the system. A parallel reader agent for Atom for instance could simply be implemented as an additional Web service. The only challenge in this regard is distinguishing the support standards of the feed sources.

- Future work will attempt to address system issues that have not been resolved at the time of this writing:

While the Library of Congress Subject Heading thesaurus is powerful, users often require a search of the personal, geographical or chronological names in news and articles. Unfortunately a free name heading has not been discovered that can be imported into the system as a supplement to, or replacement for, the LCSH.

Subjects	Result set size (number of found feeds)		
	Subject Only	Alternate Labels	Subject and Alternate Labels
Architecture	73	0	73
Bankruptcy	54	0	54
Baseball	286	1	287
Belief and doubt	0	28	28
Beverages	1	9	9
Breakfasts	3	0	3
Brunches	0	0	0
Building	150	127	277
Cell interaction	0	0	0
Cell phones	15	0	15
Charities	2	5	7
Child abuse	0	3	3
Chili powder	0	0	0
Church work	0	0	0
Civil rights	6	1	7
Clothing and dress	0	506	506
College sports	0	0	0
Composers	0	0	0
Computer security	9	5	14
Computer software-Development	0	0	0

Table 3. Result Set Size, Searching by Subject, Narrower Terms (NT) and Combined.

• The Feed Analyzer as currently implemented uses the same interval for retrieving Web feeds from all sources. This is decidedly inefficient in that the update rates of feed sources routinely differ from one another. An overly lengthy interval may result in outdated information while a short interval may waste valuable network and processing resources. Designing an algorithm to address this issue will be explored in subsequent work.

References

[1] Wang, G. A., Chen, H., Xu, J. J., Atabakhsh, H. (2006). Automatically Detecting Criminal Identity Deception: An Adaptive Detection Algorithm. *IEEE Transaction on Systems, Man, And Cybernetics-Part A: Systems and Human*, 36, 988-999, Sep.

[2] Mitchell Jr, M. B., Brown, D. E., Conklin, J. H. (2007). A Crime Forecasting Tool for the Web-Based Crime Analysis Toolkit. *in*: *Proc. SIEDS-IEEE*, p. 1-5.

[3] Madden. S., From Databases to Big Data. *IEEE Internet Computing*, 16, p. 4-6.

[4] Web Feed. Internet: http://en.wikipedia.org/wiki/Web_feed, Nov. 3, 2012 [Nov. 8, 2012].

[5] Extensible Markup Language (XML) 1.0 (Fifth Edition). Internet: http://www.w3.org/TR/REC-xml/, Nov. 26, 2008 [Nov 2, 2012].

[6] Hammersley, B. (2005). *Developing Feeds with RSS and Atom.* O'Reilly Media.

[7] Pilgrim, M. (2012). What is RSS? Internet: www.xml.com/pub/a/2002/12/18/dive-into-xml.html, Dec. 18, 2002 [Nov. 8, 2012].

[8] Ruby, S. (2008). RSS 2.0 And Atom 1.0 Compared. Internet: http://www.intertwingly.net/wiki/pie/Rss20AndAtom10Compared, Sep. 10, [Nov. 7, 2012].

[9] RSS feeds on NBCNews.com. Internet: http://www.msnbc.msn.com/id/5216556/, Feb. 21, 2012 [Nov. 8, 2012].

[10] The Official Google Reader. Internet: http://googlereader.blogspot.com/, Oct. 31, 2011 [Nov. 5, 2012].

[11] Bort, J., How Many Web Sites Are There? Internet: http://articles.businessinsider.com/2012-03-08/tech/31135231_1_websites-domain-internet, Mar. 8, 2012 [Nov. 12, 2012].

[12] Sutter, J. D. (2011). How many pages are on the

internet? Internet: http://articles.cnn.com/2011-09-12/tech/web.index_1_internet-neurons-human-brain?_s=PM:TECH, Sep. 12, [Nov. 12, 2012].

[13] Ríos, S. A., Velásquez, J. D., Yasuda, H., Aoki, T. (2006). Conceptual classification to improve a web site content.*in :IDEAL '06 Proc. of the 7th International Conference on Intelligent Data Engineering and Automated Learning*, p. 869-877.

[14] Electronic Commerce. Internet: http://id.loc.gov/authorities/sh96008434#concept, Apr. 4, 2007 [Nov. 9 2012].

[15] Library of congress Authorities and Vocabularies. Internet: http://id.loc.gov, [Nov. 8 2012].

[16] Bazargani, S., Brinkley, J., Tabrizi, N. (2013). Implementing conceptual search capability in a cloud-based feed aggregator. *In*: Presented at the 3rd IEEE Int. Conf. on Innovative Computing Technology, London, United Kingdom.

[17] Losee, R. M., (2007). Decisions in thesaurus construction and use. *Information Processing & Management* 43 (4) 958-968.

[18] SQL Server. Internet: http://www.microsoft.com/sqlserver/en/us/default.aspx, [Nov. 12, 2012].

[19] Celko, J., (2004).*Joe Celko's Trees and Hierarchies in SQL for Smarties SQL for Smarties.*

[20] Cloud computing fundamentals, *In*: *Handbook of Cloud Computing*., Furht, B., Escalante, A., Eds. US: Springer, 2010, p. 3-19. Available: http://dx.doi.org/10.1007/978-1-4419-6524-0_1.

[21] Kishore, R., Zhang, H., Ramesh, R. (2006). Enterprise integration using the agent paradigm: Foundations of multi-agent-based integrative business information systems, *Decision Support Systems*, 42, (1) 48-78, Oct.

[22] Armbrust, M., Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A. ., Stoica, I., Others. (2010). A view of cloud computing,*Communications of the ACM*, 53 (4) 50–58.

[23] Guha, R. (2010). Impact of Web 2.0 and Cloud Computing Platform on Software Engineering, in *2010 International Symposium on Electronic System Design (ISED)*, Bhubaneswar, India, p.213-218.

[24] Google App Engine. Internet: https://developers.google.com/appengine/, Jun. 26, 2012 [Jul. 27, 2012].

[25] Windows Azure. Internet: http://www.windowsazure.com/en-us/, Jul. 26, 2012 [Jul. 27, 2012].

[26] Brinkley, J., Bazargani, S., Tabrizi, N. (2012). Developing an Agent Based Feed Analyzer System in the Cloud.Presented at the 4th IEEE Int. Conf. on Cloud Computing Technology and Science, Taipei, Taiwan, 2012.