NewsSE: An Ontology-based Search Engine for News

H. Beheshti, F. Poorahangaryan, S. A. Edalatpanah*

Department of Electrical and Computer Engineering, Ayandegan Institute of Higher Education, Iran

Copyright©2017 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Rapid growth of information on the web and need to information sharing on one hand and also as news play an important role in our lives and the internet becomes the biggest repository for keeping this news on the other hand, lead us to research in this domain. In this study, we introduce a new framework for searching news by considering the relation between news and events. This framework, which is called news, considers news as a series of events in order to cover all aspects of news. News uses Domain Ontology and Event Ontology to extract the concepts and relations existed in news. Experience in using semantic techniques in search engine on the one hand and considering the relationships between events related to each news on the other hand cause an improvement in ranking results, obtaining more fresh results and cover more concepts related to a particular query compared to other search engines such as Google News and Yahoo News. The results show the Mean Average Precision (MAP) for this framework is 86.81 percent, which is higher than Google news and Yahoo! News. Furthermore the event ontology increase Average Precision about 0.8 percent compared to other traditional search engines.

Keywords Semantic Search Engine, Inference Engine, News Search Engine, Ontology-based Search Engine

1. Introduction

The creation of the World Wide Web (WWW) was a major step toward information sharing. It resulted in the development of techniques which make information sharing easier and more powerful. Information Retrieval is one of the main tools to reach this goal. In the recent years, as volume of information on the Web increase rapidly, use of search engines for exploring the Internet becomes more urgent. This information explosion causes finding correct and relevant contents in the Web become more complex. These shortcomings clarify the need to use new approaches and technologies for improving the performance of searching information on the Web. One of these technologies is Semantic Web.

Nomenclatures	
t	Current time
i	Number of news resources
T	Local time of news resources
P	Time portion
δ	Time interval related to P
Q	Set of terms
S	Set of synonyms related to each q_i
D	Set of related documents to each q_i and s_i
t_i	Keyword number i
d_k	Document number k
a_{ik}	Number of the keyword t_i in a document d_k
f_{ik}	Term frequency of a term t_i in the document d_k
n_i	Number of documents that contains keyword t_i
N	Total number of documents in the corpus
Greek Symbols	
α	Degree of freshness.
β	Degree of credibility
λ	Query selection threshold
Abbreviations	
RDF	Resource Description Framework
IR	Information Retrieval
DO	Domain Ontology
STI	Static Time Interval
XML	Extensible Mark-up Language
TRAD	Traditional Search Engine (without Ontology and Inference Engine)
CB	Ontology-based Search Engine (with Domain Ontology)
EB	Event-based Semantic Search Engine (NewsSE)
MAP	Mean Average Precision

New generation of search engines use some Semantic Web tools such as RDF, Ontology, Annotation and Inference Engine in order to present the most relevant results to the users. In fact the goal of using these techniques is to approach generated results to user's queries.

We can divide existing data on the Web to different categories. One of these categories is news. Nowadays, The Internet is the biggest resource for keeping news. This huge amount of news in the Internet indicates the necessity of using powerful Information Retrieval mechanisms in order to search among them precisely.

Today there are many search engines which still use traditional methods for searching information on the Web. These engines use different methods and techniques. These traditional (keyword-based) search engines work based on the General IR model. One of the most popular search engines which is well known as the best search engine called Google [1]. This company achieved better results for many searches with an innovation called PageRank. This iterative algorithm ranks Web pages based on the number and

PageRank of other Web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Google also maintained a simple and user friendly interface to its search engine. In contrast, many of its competitors embedded a search engine in a Web portal.

News browsing and searching is one of the most important Internet activities. The huge amount of news available online reflects the users' need for a plurality of information and opinions. News Search engines are a direct link to fresh and unfiltered stream of information. There are many commercial News search engines such as Google News retrieves news information from more than 4,000 sources, organizes it into categories and automatically builds a Web page with the most important news for each category. Yahoo News runs an analogous service on more than 5,000 sources. Microsoft recently announced its NewsBot, a news engine that provides personalized news, according to different profiles built for each user [2].

As the importance of the news search engines, there are numerous researches by the scholars. NewsInEssence [3] is a

system for finding and summarizing clusters of related news articles. QCS [4] is a software tool for streamlined IR from generic document sets. [5] Proposes a tool to automatically extracting news from Web sites. NewsJunkie[6] is a system that personalizes news for users by identifying the novelty of stories in the context of stories that users have already reviewed. SmartWeb [7] is another effort in order to develop a number of domain-specific ontologies that are relevant for mobile and intelligent user interfaces to open-domain question-answering and information services on the Web.

According to previous research, none of the news search engines care about the relationship between events and news. The objective of this paper is achieving a high retrieval performance by using two different kinds of ontologies; Domain ontology and Event Ontology. This study presents a complete Ontology-based framework for the extraction and retrieval of semantic information in the limited domain of news by considering events. We applied the framework in football domain and observed the improvements over traditional keyword-based approaches. This research uses Event Ontology and Domain Ontology as two main building blocks of search framework in order to increase the result coverage. The outline of this paper is as follows. In the next section, a NewsSE framework is presented. Experimental results are discussed in Section III. Finally, conclusions are drawn in Section IV.

2. Proposed Method (NewsSE)

Although various search engines work based on various techniques, but almost none of them don't care about the structure of information which should be searched. Moreover, the majority of search engines which work on the news don't customize for searching news. Therefore they suffer from low precision and performance.

In this study, we introduce a new framework which is used especially for searching in the news. This framework consists of several processes. Moreover, we focus on events that are related to news and then extract relation between them in order to increase the performance. In all processes of this framework, we use semantic techniques. The descriptions of the processes are shown in Table 1.

In the following sections we explain each of these processes. Figure 1 shows NewsSE framework and its processes. In fact, by using Semantic Web methods, we want to substitute attention to the aspect (appearance) of queries with attention to the meaning of queries.

2.1. Ontology

Ontology keeps all concepts and the relation between them. Moreover, they play a central role in Semantic Web applications by providing a shared knowledge about the objects in the real world, which promotes reusability and interoperability among different modules. Therefore, the quality of the Ontology should be the first concern in any semantic application.

Ontology that is created and used in this study have been evaluated based on human judgment. If we can prove the positive effects of using these two limited ontologies in this research, we can generalize the results to other ontologies and domains.

Since covering all different types of news is very difficult, this study focuses on a limited scope. We choose the football domain for this reason. Therefore, in the rest of this research all examples around the news are related to football.

NewsSE uses two different ontologies. First Ontology is called Domain Ontology and another one is called Event Ontology. In the following parts, we explain each of these ontologies.

2.1.1. Domain Ontology

This Ontology contains all concepts, relations and instances (individuals) related to a specific domain (football domain). For creating this Ontology we use protégé [8] Figure 2 shows the body of this Ontology.

Football Ontology (FO) contains 31 classes and 26 properties. The full class hierarchy can be seen in Fig. 3. This Figure shows only hasSubclass property.

Indexing and query enrichment processes use this Ontology. In fact, the existing rules in the inference engine must be run over this Ontology in order to lead to enriching the query and indexing process.

2.1.2. Event Ontolog

One of the major goals of this research is focusing on events and relations between them. In this research, we design Ontology for extracting this event. This Ontology called Event Ontology. The Figure 4 shows the body of this Ontology. It should be noted in this Ontology, we just focus on five important events in football domain.

Here, different relationships are available between all classes in the Event Ontology. Figure 5 shows Event Ontology and its relationship with Domain Ontology. Event Ontology is used similar to Domain Ontology.

Table 1. The descriptions of the NewsSE processes

Process	Description
NewsSE	A new framework for searching in news based on events.
NewsCr	A crawler which uses new methodology for decreasing the number of crawls
NewsDex	An indexer which uses new algorithm for indexing by focusing on concepts and events related to news
NewsRank	Introduces a new ranking algorithm (NewsRank) which is used just for ranking news
NewsQ	A query processor which uses new algorithm for enriching user's queries. The heart of this new algorithm is a rule-based Inference Engine

Figure 1. NewsSE Framework

Figure 2. Football Ontology in Protégé Editor

Figure 3. Class Hierarchy of Football Ontology

Figure 4. Football Event Ontology

Event Ontology

Figure 5. Relation between Domain Ontology and Event Ontology

Figure 6. Process of Determining the Crawling Time Interval

2.2. NewsCr

Retrieving news from the Web and giving them to the Indexer module are the main goal of each crawler. In fact, a crawler is a Web application which works such as a robot. Each crawler contains a list of addresses (URLs). Addresses point to the news resources. Crawler surfs the Web in certain interval and fetch fresh news in different formats and store them in a temporal repository. Our crawler can fetch news in XML, RSS [9] and HTML formats.

Today, determining the time interval between two successive crawls to the same resource is a major concern and the lack of attention to these causes reducing efficiency of search engine and wasting system resources. Therefore, in NewsCr, we introduce a dynamic methodology in order to specify the time interval between two sequential crawls for a specific news resource. If we select a static time interval between two sequential crawls, it is possible some crawls that get ineffective. Figure 6 shows our new methodology for determining the time interval.

As shown in Figure 6, t indicates our current time (crawler time), i is the number of news, resources, T specifies the local time of news resources, P indicates the time portion and delta is the time interval related to specify P (Table 2). The first step of this process is calculating the time difference between crawler current time and news resources time. Second step is determining the appropriate time portion for news resources and finally allocating time interval related to specified portion to the crawler scheduler.

Table 2. Time Interval related to each Portion

Portion (P)	Time range	Time interval (delta)
1	6 – 16	5 minutes
2	16 – 22	10 minutes
3	22 – 6	30 minutes

To find an appropriate time interval for each portion, we monitor the rate of arrival news into several News Agencies (resources) and indicate peak times. Based on this finding, we divide time into three portions and then we allocate time interval (delta) to each resource.

For example, suppose in a country like Iran, the peak of arrival news into News Agencies is between 6 am until 4 pm. Moreover, suppose we select a Static Time Interval (STI) equal to 20 minutes. In these conditions we can see a big gap between two sequential crawls in peak time. In contrast to this condition, in the midnight we have a lot of unnecessary crawls. These situations decrease the performance of the system and also cause wasting a lot of resources.

2.3. NewsQ

Users enter their queries via the User Interface (UI). Query Enrichment is a process used in order to make user's queries more enrich. In this study, we introduce a new methodology for enriching user's query called NewsQ. NewsQ contains several sub processes which are described in the Fig. 7.

Figure 7. Process of Query Enrichment

In the following, we explain each of the phases shown in

the Figure 7:

- In the first two steps, we want to create sets of terms by tokenizing and eliminating all *stopwords*. The output of this part is a set which contains several terms ($Q=\{q_1, q_2, \dots, q_{ij}\}$)

By using *WordNet*, we should find all synonyms of terms and create the set of synonyms related to each q_i . The output of this process is a set called $S=\{s_1, s_2, \dots, s_{ij}\}$.

- In this section, type of q_i and s_i should be realized. These types include:
 - *Class names*: all classes in the Domain Ontology and Event Ontology
 - *Individuals*: all instances of classes in Ontologies (Knowledge base)
- After determining the type of each q_i and s_i , the system allocates the appropriate weight to each of them. This process is controlled and executed by rules in the Inference Engine. (In the next section, rules and inference engine will be explained)
- In this stage, we define D as a set of related documents to each q_i and s_i and then we calculate sum of weights related to each document.
- At the final stage, we choose an appropriate threshold. All documents with weight more than this threshold is our results.

For better understanding this process, we mention an example in Fig. 8.

The output of Query Enrichment process is a set of unordered results with their weight. The ranking process uses this weight to order results.

2.4. Inference and Rules

The rule and inference engine is used in indexing process and Query Enrichment. Each inference engine consists of a set of rules that if a certain condition is satisfied they should be fired. These rules help the system to extract new knowledge from Ontology and Knowledge-base. For example Query Enrichment uses rules to term weighting or indexing process uses rules to extract event from news and relationship between the concepts. Following list shows several rules used in NewsQ.

Figure 8. An Example of Query Enrichment

- If q_i is a Class name then $\lambda=2$
- If q_i is an Individual then $\lambda=3$
- If q_i is a specific Class then $\lambda=3$
- If q_i is a specific Individual then $\lambda=4$

Inference is a costly process, especially if the number of assertions is large. This problem should be handled carefully to maintain the scalability of the system. In evaluation section we will be familiar with set of rules.

2.5. NewsDex

Indexing is one of the important processes of any retrieval system. The important technology used for indexing is “term weighting”. The value (denoted as a_{ik}) of the element of a keyword t_i in a document d_k is determined by different term weighting algorithms. In this part, first we should be familiar with some expressions. In this research we denote f_{ik} to be the term frequency of a term t_i in the document d_k (news), which means the number of occurrence of term t_i in the document d_k divided by the total number of words in the document d_k . We denote n_i to be the number of documents that contains keyword t_i . We denote N as the total number of documents in

the corpus. So $\frac{n_i}{N}$ means the percentage of documents containing term t_i among all documents in the corpus, which we call document frequency.

NewsDex uses the Boolean weighting method. This is a simple method. It assigns the weight to be 1 if the word appears in the document (news), 0 if not.

Eq. (1) shows how this method calculates the weight:

$$\begin{cases} 1 & \text{if } f_{ik} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

For indexing news, we use this method because of finding term frequency in a statement is not reasonable [10]. We just need to know a word appear in a news title or not. As we would like to incorporate into indexing process, we add some features to this method in order to improve the process of indexing. Figure 9 shows the overall process of NewsDex.

As shown in Fig. 9, the process of indexing contains 4 stages:

1. Tokenization

In this stage, each news should be tokenized into several phrases. For performing tokenization we use a parser.

2. Eliminating Stopword

In each search engine, a Stopword is a commonly used word (such as "the") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. When building the index, most engines are programmed to remove certain words from any index entry. The list of words that are not to be added is called a stop list. Stopwords are deemed irrelevant for searching purposes because they occur frequently in the language for which the indexing engine has been tuned. So in this stage we should remove all Stopword by comparing each terms of news with list of Stopword.

3. Calculating the Weight of Terms, Concepts and Events in Each News

For calculating the weight, we use Boolean Method for each news. In addition to the weighting terms in each news, we weight concepts and events related to news. For finding all concepts and events related to news, we have to use Domain Ontology and Event Ontology respectively.

4. Index Matrix

After each indexing process, the index matrix should be updated. Table 3 shows a forward index matrix. Each row of matrix shown in the Table 3 indicates a specific document and each column specifies a term or concept or event.

Figure 9. Indexing Process of a Typical Document

Table 3. Forward Index Matrix

	T1	T2	.	.	.	Ti	C1	C2	.	.	.	Cj	E1	E2	.	.	.	EK
news 1	0	0	.	.	.	1	1	0	.	.	.	1	0	0	.	.	.	0
news 2	1	0	.	.	.	1	1	0	.	.	.	0	0	0	.	.	.	1
news 3	1	0	.	.	.	0	0	1	.	.	.	0	0	0	.	.	.	0
news 4	0	0	.	.	.	0	1	0	.	.	.	1	1	0	.	.	.	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
news n	0	1	.	.	.	0	0	1	.	.	.	0	0	1	.	.	.	1

5. Storing Indices in XML Format

The structure of the semantic index has high importance in the retrieval performance. We constructed a XML (Extensible Mark-up Language) format for storing indices of each news. Table 4 shows an example for index structure of a specific news.

Table 4. Index Structure

News Id	12
Title	Messi extend his contract with Barcelona for 3 years.
Term indices	Messi - contract – Barcelona
Concept indices	Football - Lionel Messi - Argentina National Team - ...
Event indices	Transfer
date	2015-03-05
Time	16:30
Credibility	4

2.6. NewsDex

All the news that are chosen as a result in the previous steps, at this stage must be ranked in order to present in the output. Therefore we present new algorithm for ranking news called NewsRank. In this study we assume that there are two different preferences for ordering the results:

- *News freshness*: if users select this item, all results ordered just by NewsRank algorithm.
- *Similarity with user's query*: if users select this item, before using NewsRank, another process should be done. In this process, a threshold should be chosen and all news with overall weight more than this threshold should be ordered by NewsRank algorithm.

All processes done in a search engine can be presented as well, only if search engine uses a suitable ranking algorithm. If all processes are executed well but the results are not properly presented to users, all efforts are wasted. After finishing all tasks, generated result should be ordered based on specific algorithm. Figure 10 shows NewsRank algorithm.

Figure 10. NewsRank Algorithm

Since ordering and ranking process in a news search engine is different from other kinds of search engines, we choose two important parameters for each news; time of occurring news (or degree of freshness) and news resource (or degree of credibility). In the above algorithm, we denoted these two parameters with α and β . The algorithm works as follows:

- First, all generated results should be ordered base on the parameter α (degree of freshness). At the end of this stage each news with smallest α should be presented in higher position.
- In this stage, news with equal α should be reordered based on second parameter β .

This mechanism ensures more recent and more reliable news are placed in the higher position. Specifying NewsRank parameters is a manual process. In fact, reliable resources have bigger β . Using β parameter helps to the engine to control the process of showing results. For example, it's possible in some situation we change the β of a resource in order to increase or decrease the importance of a resource. If we decide not to show the results of a specific resource, we can set β equal to zero.

3. Results and Discussions

In this section, we assess and compare the experimental results and proposed framework with two famous news search engines; Google news and Yahoo! News. Before starting the section, it is important to note that both Google and Yahoo search engines in terms of store size are very powerful and also as accessing to resources stored in these two engines are not possible therefore we focus just on first 50 results of each query. In the following parts, first we introduce set of queries used for evaluation, then we evaluate NewsSE performance.

In order to evaluate the proposed system, 20000 sports news has collected from various sources.

3.1. Evaluation Queries

We started the evaluation process by preparing the 8 queries shown in Table 5. In this research we classified the queries into 8 different classes that each query play the role of an indicator for each class. Next to each query, we also put the corresponding keyword query which was actually used in the evaluation.

Table 5. List of Queries

	Query	Keywords
Q1	Find all goals	Goals
Q2	Goals scored by Barcelona	Goals - Barcelona
Q3	Goals scored by Messi at Barcelona	Goals – Messi - Barcelona
Q4	Punishments	Punishment
Q5	Yellow cards received by Alex	Yellow card - Alex
Q6	punishment of Henry	Punishment - Henry
Q7	Which team Ronaldo transfer to	Ronaldo – Team - Transfer
Q8	Goalkeeper of Barcelona	Goalkeeper - Barcelona

3.2. Overall Evaluation

Table 6 presents the Mean Average Precision (MAP) [11, 12, 13] for three different type of search engines.

Table 6. Evaluation Result (Mean Average Precision)

	TRAD	CB	EB
Q1	57.96%	84.91%	81.06%
Q2	51.43%	78.26%	84.26%
Q3	41.03%	83.88%	79.28%
Q4	59.19%	81.32%	94.37%
Q5	9.20%	84.07%	89.54%
Q6	17.72%	62.52%	95.37%
Q7	26.05%	66.90%	83.12%
Q8	7.26%	84.90%	87.51%
Average	33.73%	78.35%	86.81%

Figure 11 shows the evaluation result for comparing MAP of TRAD, CB and NewsSE. As shown in Figure 11, we have different situation for each query. In first three queries, we don't have considerable differences between CB and EB because the framework cannot extract any event from these queries.

Figure 11. Evaluation Results for Comparing TRAD, CB and NewsSE

For Queries n, Q4, Q5, Q6 and Q7, we can see the effect of adding Event Ontology to this to this framework. As Event Ontology helps us to extent the coverage of results so it shows better AP. The last query presents the effect of Inference Engine because in this query the name of Goalkeeper is not given explicitly but this system can infer the name of them from the Domain Ontology. At the final stage, we compare our framework with Google news and Yahoo news. The Table 7 shows the MAP for these three engines.

Table 7. Comparing our Framework with Google and Yahoo

	Yahoo News	Google News	NewsSE
Mean Average Precision (MAP)	78.97%	86.01%	86.81%

4. Conclusions

In this research, we presented a complete framework for performing semantic search in news. In addition to focus on the news in this framework, we considered events as the major components of news. In fact, in addition to the use of Semantic Web techniques that enhance the efficiency of search operations, the particular attention to news and events have a large impact on improving the quality of generated results for each query. Because of above mentioned reasons, the Mean Average Precision (MAP) for proposed framework is 86.81 percent which is higher than Google news and Yahoo! News. Furthermore the event ontology increase Average Precision about 0.8 percent compared to other traditional search engines.

REFERENCES

- [1] Brin, S; and Page, L. (2012). The anatomy of a large-scale hypertextual web search engine. *Journal of Computer Network*, 56(18), 3825-3833.
- [2] Gulli, A. (2005). The anatomy of a news search engine. *Special interest tracks and posters of the 14th international conference on World Wide Web*. ACM, 880-881.
- [3] Radev, D. R., et al. (2001). Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. *Proceedings of the first international conference on Human language technology research*. PA, USA. 1-4.
- [4] Dunlavy, D. M., et al. (2003). QCS: a tool for querying, clustering, and summarizing documents. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Edmonton, Canada. 11-12.
- [5] Reis, D. d. C., et al. (2004). Automatic web news extraction using tree edit distance. *Proceedings of the 13th international conference on World Wide Web*. New York, USA. 502-511.
- [6] Gabrilovich, E., et al. (2004). Newsjunkie: providing personalized newsfeeds via analysis of information novelty. *Proceedings of the 13th international conference on World Wide Web*. New York, USA. 482-490.
- [7] Oberle, D., et al. (in press). Dolce ergo sumo: On foundational and domain models in swinto (smartweb integrated ontology). *Journal of Web Semantics*.
- [8] Ontology, P. (2007). Knowledge Acquisition System. from <http://protege.stanford.edu>.
- [9] Guha, R., et al. (2003). Semantic search. *Proceedings of the 12th international conference on World Wide Web*. New York, USA. 700-709.
- [10] Madani, O. and J. Yu (2010). Discovery of numerous specific topics via term co-occurrence analysis. *Proceedings of the 19th ACM international conference on Information and knowledge management*. New York. USA. 1841-1844.
- [11] Tumer, D., et al. (2009). *An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, Yahoo, MSN and Hakia*. Internet Monitoring and Protection, 2009. ICIMP'09. Fourth International Conference on, IEEE. 51-55.
- [12] Büttcher, S., et al. (2010). *Information retrieval: Implementing and evaluating search engines*. Mit Press.
- [13] Lee, D., et al. (2007). *Improvement of the recall and the precision for semantic web services search*. Computer and Information Science. ICIS 2007. 6th IEEE/ACIS International Conference on, IEEE. 763-768.