Metrics for a Semantic Search Engine for the Sensor Web

William Herbert
Oakland University
6632 Telegraph Rd, #325
Bloomfield Hills, MI 48301
248-703-9159
wgherber@oakland.edu

Fatma Mili
Oakland University
157 Dodge Hall of Engineering
Rochester, MI 48309-4401
248-370-2246
mili@oakland.edu

Yahia Khelifa
Oakland University
157 Dodge Hall of Engineering
Rochester, MI 48309-4401
248-370-2246
yjkhelif@oakland.edu

ABSTRACT
Traditionally, web searching has focused primarily on web pages. Mechanisms for finding, indexing, and ranking these pages have involved linking among pages. However, the overwhelming majority of web information is not stored in web pages. This content, known as the Deep Web, exists in such forms as images and databases (including sensor network data). Because Deep Web resources are not "connected" or self describing in the same way as are surface resources, they cannot be searched and accessed using traditional surface web search methods. In this paper we focus on a particular type of Deep Web resource, the web-enabled Sensor Network, or Sensor Web (SW). We discuss the challenges in locating and identifying the contents of Sensor Webs and we discuss metrics needed to assess the relevance (actual and potential) and authority of SWs, individually and in groups, with respect to an information request or query. We define and discuss a 3 dimensional vector of metrics (topicality, coverage and timing) that quantifies the degree to which a SW addresses the *what, where, and when (also how often)* aspects of a query. We use semantic distance to measure the topicality relevance of a SW to a set of query terms. We use geometric overlap metrics to measure coverage relevance; we use time, time range and frequency of observation metrics to measure timing relevance. To quantify the authority of a SW, we use a variation of topic-sensitive PageRank to assess the importance of the entity responsible for the creation and maintenance of the SW resource. Such entities are usually identifiable via the surface web. We present examples of the use of these metrics and describe a project to further investigate their use in facilitating effective deep web searching and the development of a search engine for Sensor Web resources. We validate our approach in a two step process that will involve: 1) Creation of a computer model of the proposed metrics and ranking methodology, and 2) Testing of the model with a predefined set of queries and SW resources by a panel of test users who will assess and evaluate the rankings of the SW resources returned by the model with respect to various queries. Model parameters will be adjusted based on test feedback in order to evaluate and optimize the ranking approach.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Relevance feedback, Retrieval models, Selection process*

General Terms
Measurement, Algorithms

Keywords
Semantic Web, semantic distance, sensor web, relevance, authority

1. INTRODUCTION
"Searching on the Internet today can be compared to dragging a net across the surface of the ocean. While a great deal may be caught in the net, there is still a wealth of information that is deep, and therefore, missed. The reason is simple: Most of the Web's information is buried far down on dynamically generated sites, and standard search engines never find it." *Michael K. Bergman*

Search engines have been developed and refined with a focus on locating web pages, i.e. primarily static documents. Web pages have many features that have been exploited by search engines to find them, index them, and rank them. The uniform header information as well as the machine readable text content is used to determine the nature of the information contained in these pages. The links across pages are used by crawlers to locate new pages and to assess standing of a page among its "peers." As useful as these pages and search engines are, they can be seen as just the surface of a very deep ocean of information. The phrase "Deep Web" was coined by Michael Bergman [1] to refer to the wealth of information that is located in repositories accessible on the internet, yet remain largely invisible to traditional search engines.

In 2001, Bergman estimated the deep web content at nearly 550 billion individual documents compared to one billion of the surface web. He observed that the search engines with the largest number of web pages indexed at that time, such as Google or Northern Light, each indexed no more than 16% of the surface web. By missing the deep web, Internet searchers using these engines were searching only 0.03% of the available information content of the web at the time. A 2009 survey by He, et. al. [2] updates those numbers, measuring the size of the deep web in terms of deep-web sites, back-end web databases and query interfaces, where a deep-web site is a web server that provides information maintained in one or more back-end web databases, each of which is searchable through one or more HTML forms as its query interfaces. He estimates there to be 307,000 deep web sites, 450,000 web databases (of which 348,000 are structured and 102,000 unstructured (e.g. texts, images, audio, and video)). He further estimates that the deep web has expanded 3 - 7 times in the 4 years (2000 - 2004). He also observes that the three search engines, Google, Yahoo and MSN index respectively 32%, 32% and 11% of the deep web. But because there is considerable overlap in their coverage the overall indexing is only 37%. Thus,

while the deep web is not quite as hidden as it was a few years ago, the majority of it still cannot be accessed by current search methods.

An important web resource that is currently not on the radar of search engines is the data generated by sensor networks (SNs) which are computer accessible networks of many spatially distributed devices using sensors to monitor conditions at different locations, such as temperature, sound, vibration, pressure, motion or pollutants [3]. SNs are a component of the structured databases component of the deep web as referenced above. In recent years, SNs, particularly wireless sensor network (WSNs) have seen explosive growth in the number of their applications, and have become ubiquitous in modern society. As a major growth technology, WSN R&D budgets are projected to rise to $1.3 billion in 2012, up from $522 million in 2007, according to a recent study by ON World [4].

SNs that are web accessible are referred to as sensor webs (SWs). A SW can encompass current observations or archived data and can consist of a single SN or a set of SNs that are able to communicate with each other via the web (referred to in [5] as a "web of webs"). Within this framework, the term sensor web can refer to a single web-enabled sensor network or to the collection of all sensor webs. Because they are currently not accessible by search engines, SWs are a part of the deep web and can be considered a subset of the databases category since, like databases, they generally are accessed only by query.

The rapid growth in the deployment of sensor networks, coupled with the lack of standard practices and protocols for their use has resulted in a lack of communication, integration and interoperability among sensor networks. The Open Geospatial Consortium (OGC; www.opengeospatial.org), an international consortium of companies, government agencies and universities, was formed in 1994 to address this situation. Similar to the manner in which HTML and HTTP standards enabled the exchange of information on the Web, the OGC's Sensor Web Enablement (SWE) initiative is focused on developing standards to enable the discovery, exchange, and processing of sensor observations, as well as the tasking of sensor systems [6]. OGC has promoted sensor web enablement through the establishment of several encodings [7] for describing sensors and sensor observations, and through several standard web service interface definitions for the discovery and tasking of sensor networks.

The World Wide Web consortium (W3C; www.w3.org/2001/sw/) has also played a major role in the SWE effort by promoting standardization through its Semantic Web Activity. This activity provides a common framework that allows data to be shared and reused across application, enterprise and community boundaries. W3C's Semantic Web is based on the Resource Description Framework (RDF), a W3C standard model for data interchange on the Web [8]. RDF extends the linking structure of the Web by using URIs to identify the relationship between things (usually referred to as a subject/predicate/object triple).

2. CHALLENGES AND OPPORTUNITIES IN SEARCHING THE SENSOR WEB

The literature on the deep web [9, 10] outlines many of the main challenges of locating, assessing, and ranking resources below the surface. The main challenges are: 1). Deep web resources are not inherently "connected" in the same way that surface resources are and thus require a different mechanism to locate them. Surface Web resources such as HTML pages are connected to other Web resources by internal URL references within a page to another page. Deep Web resources have no similar connectedness. 2). Deep web resources are not self describing in the same way that surface resources are. Unlike surface Web resources, such as HTML pages that contain descriptive text, narratives, charts and graphs that can easily be interpreted, deep web content, such as databases and WSNs, is usually opaque and cannot be understood without the benefit of a catalog, index or some other interpretation tool/process. 3). In part as a result of the first two points, deep web resources require a different mechanism for assessing their relevance and authority. Because of the self describing nature of surface web pages and their interconnectedness through URL references, relevance of such pages can be quantified in terms of the number of occurrences and the physical location of keywords. Similarly, the authority of a page can be measured in terms of the number of times it is referred to by other web pages. However, because deep Web resources are not self descriptive or linked, alternative methods are needed to determine relevance and authority that do not require an examination and analysis of the contents of the resource. Sensor networks share some of these same issues. In particular:

Difficulty in locating. Sensor networks are usually created to be self contained and independent and generally do not link to other networks. Generally this is addressed (at least in the short term) by requiring participating SWs to sign up in a registry.

Difficulty in identifying the "contents" of the data available from sensor networks. Evolving SWE standard protocols for describing and annotating sensor networks such as Sensor Modeling Language and standard metadata encodings are helping to improve this situation, however as these standards are refined, new procedures should not place too much of a burden on the SN owner.

Need for a different metric to assess the relevance of a sensor web to a query. Relevance between a query and an SW is inversely proportional to the semantic distance between the query and the content of the SW. On the one hand this issue is challenging with SWs because of the lack of an overhead-free mechanism of identifying contents, on the other hand it is made somewhat easier because SWs have a set of features that can be easily standardized: data collected, time stamps, geographic location, accuracy, etc.

Need for a metric to assess the relevance of a group of sensor webs. The combination of two networks may have a high relevance factor even though neither of the individual networks has high relevance. Metrics need to be developed for compositions of SWs.

Need for a new metric to assess authority. Whereas surface web resources use links to assess authority and trustworthiness, we need a different mechanism for SWs, e.g. the authority and trustworthiness of the owning entity combined with some additional factors.

Need to distinguish between actual relevance and potential relevance. A SW may not have a high relevance factor as is, but may gain a higher relevance factor if slightly modified (nodes

asked to move to area of interest, or to change their frequency of sampling, or the nature of the sampling).

In this paper we focus on issues of relevance and authority. In section 3 we discuss metrics for quantifying the relevance of a SW to a query; in section 4 we discuss authority metrics for a SW relative to a query; in section 5 we discuss the experimental design as well as future plans to generalize these metrics to a composition of SWs. We summarize and conclude in section 6.

3. RELEVANCE METRICS

The difference between querying a specific SW and sending a request by searching the deep web is illustrated in Figures 1 and 2. In the targeted scenario, the query is a "native" expression formulated precisely in terms of the capabilities and functionality of the SW in question. By contrast, in web searching, the request is a high level desired result that needs to be mapped to existing resources and capabilities and matched, then a formal expression is constructed.

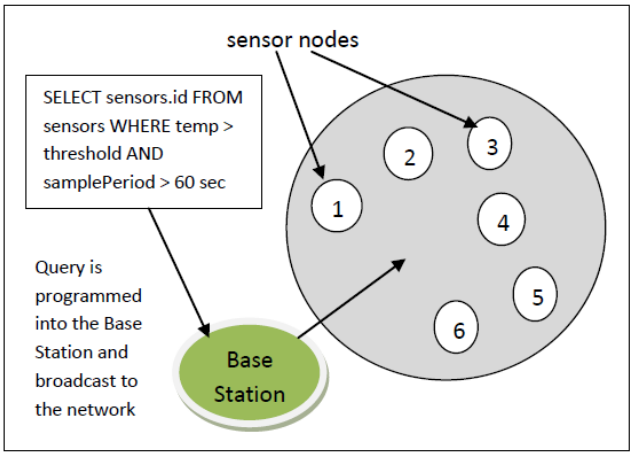

Figure 1: Targeted querying of a specific SW

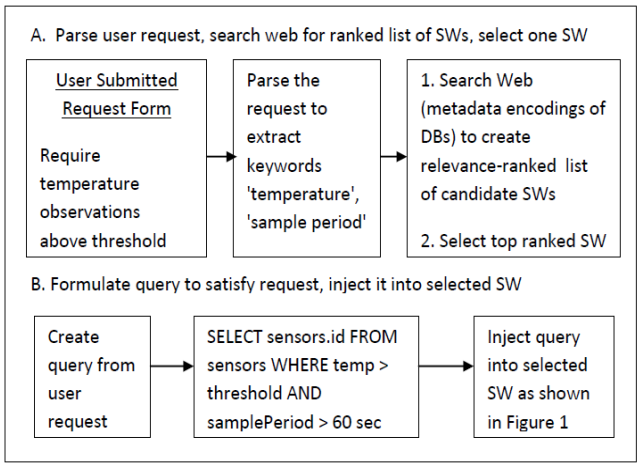

Figure 2: Two stage process for searching and querying the web for SW data

The examples in Figure 3 further illustrate the difference. 3.1 Shows a formal query to a specific SW. 3.2 and 3.3 show the components of a high level user request to identify and search for a sensor web that meets certain criteria. When the most relevant and authoritative source is found, a query is formulated similar to 2.B and injected in to the chosen SW.

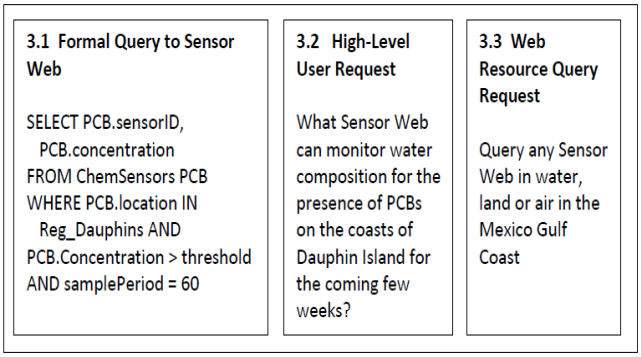

Figure 3: Targeted querying vs. Information request, search and query

To distinguish between a formal query custom-tailored to a SW and a less specific search, we call the latter a *request* rather than a query. Irrespective of its level of formality or specificity, any query or request can be characterized by three parameters: *what, where, and when (also how often).* Similarly, every SW can also be characterized by these same parameters, *what* is the nature of the data that the network can sense (e.g. barometric pressure, dew point, etc. within a domain of interest such as weather conditions); *where* is the sensor network located, (e.g. latitude, longitude, altitude, distance below earth's surface, etc.) as well as the granularity of observation (e.g. number of observation per square meter); and *when* are the measurements taken. This addresses frequency and any other time constraints. We call these three parameters *topicality*, *coverage* and *timing*.

Relevance of a SW to a request is a 3 dimensional vector of metrics that quantifies the degree to which a SW addresses the topicality, coverage and timing requirements of a request or query. We discuss each of the three metrics in turn.

3.1 Topicality Metric: Measuring the Topicality Relevance of a SW to an information request

The topicality of a SW specifies the subject matter (i.e. the *what* parameter) of a domain within which the SW can make observations. Given a SW and a user request for data about a specific topic, we want to be able to assess the degree to which the SW satisfies the topicality requirements of the request. This is the topicality relevance of the SW to the request. If the topic of a SW observation exactly matches the topic of a query term, we say that the SW has full topical relevance to the query and define the semantic distance between the two terms as 0. Conversely, If the topics are completely dissimilar, we say the SW has no topical relevance to the query and define the semantic distance between as 1. We calculate semantic distance using the semantic distance method described by Corby in [11] and an ontology-based topic

description such as the *Semantic Web for Earth and Environmental Terminology* (SWEET) ontology [12]. We normalize the resulting semantic distances to the interval [0, 1]. This allows us to use topical relevance (defined by semantic distance) as a vector element in a similarity vector that quantifies the overall similarity of a SW to a query. In addition to topical relevance, the similarity vector will include other elements, all normalized to [0, 1], such as coverage relevance, timing relevance, authority and other parameters described in this paper.

Corby expresses semantic distance within the framework of a directed graph in which nodes are ontological terms and edges are relations among the nodes. He defines the semantic distance between two nodes as the sum of the lengths of the shortest paths between each of them and a common parent. He further states that, because, low level ontology classes are semantically closer than top level classes, the ontological distance between nodes decreases as their depth increases. Thus, in the segment of the SWEET ontology shown in the tree of Figure 4, w*etland* and *lake*, which are brothers at depth 5, are closer than *property* and *realm* at depth 1.

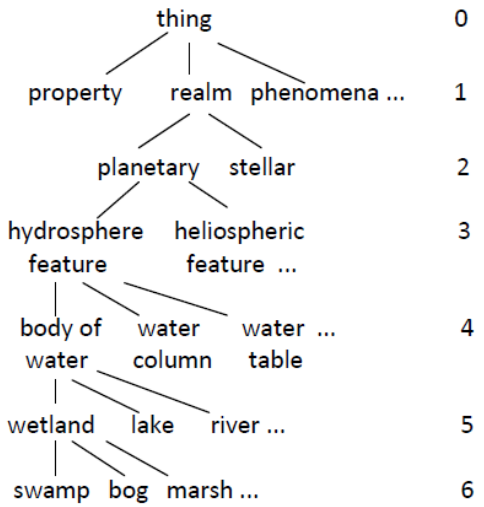

Figure 4: Partial tree representation of the SWEET Ontology

As an example, suppose R_1 is a request for a measurement of the surface temperature of a swamp at a certain geographic location. For R_1, there are two topical query elements; *swamp* and *temperature*. Suppose, further that there are three environmental monitoring SWs available to take water temperature observations at the location of interest. Suppose, finally, that none of the three SWs have sensing capabilities specifically targeted to swamps. SW_1 takes observations in wetlands, lakes and rivers, SW_2 takes observations in lakes and rivers, while SW_3 takes readings in bodies of water and water columns. Although none of the three candidate SW exactly addresses the topic *swamp*, perhaps one of them makes observations of a class that is close enough to swamp, that R_1 can be reasonably well satisfied. Thus, we want to know which of the three candidate SWs has the highest topical relevance to R_1. (We are only concerned about the relevance of the three SWs with respect to the query term *swamp*. All three SWs are fully relevant with respect to the query term *temperature*, since they all observe temperature).

Corby defines the length of link between a node t and its direct super type (parent) t' in an inheritance hierarchy *H* by $\frac{1}{2^{d_H(t')}}$ where $d_H(t')$ is the depth of t' in *H*. He further defines the distance between two nodes as the minimum sum of the lengths of the paths between each of them and a common super type. Referring to Figure 4, the semantic distance, d(swamp, lake), between the query term *swamp* and the observation term *lake* is defined as follows:

d(swamp, lake) = d(swamp, wetland) + d(wetland, body of water) + d(lake, body of water)

$$= \frac{1}{2^5} + \frac{1}{2^4} + \frac{1}{2^4} = \frac{5}{32}$$

Similarly, d(swamp, river) = $\frac{5}{32}$, d(swamp, wetland) = $\frac{1}{32}$, d(swamp, body of water) = $\frac{3}{32}$ and d(swamp, water column) = $\frac{11}{32}$. We say that d(swamp, swamp) = 0.

If wetland is at the lowest depth (6) of the tree, then the distance from the term *swamp* to the farthest node in the graph $\leq \frac{63}{64}$. Normalizing all semantic distances to this maximum distance results in the distances as shown in Table 1.

Table 1: Topical Relevance of 3 SWs to query term *swamp* of R_1

SW ID	Observation Term	Normalized Semantic Distance from Query	Ranking of SWs with respect to Topical Relevance
SW_1	**wetland**	**0.032**	1
	river	0.159	
	lake	0.159	
SW_2	**river**	**0.159**	3
	lake	**0.159**	
SW_3	**body of water**	**0.095**	2
	water column	0.338	

3.2 Coverage Metric: Measuring the Coverage Relevance of a SW to an information request

The coverage of a SW specifies the geographic area over which the SW can make observations. Given a SW and a user request for data about a specific location, we want to be able to assess the degree to which the SW satisfies the coverage requirements of the request. This is the coverage relevance of the SW to the request. We use geometric overlap metrics to measure coverage relevance.

We define coverage overlap error COE as follows. If CR is the coverage required by the user request, and CA is the portion of CR that is actually covered by the SW, then:

$$COE = 1.0 - CA/CR$$

Thus, If the geographic area sensed by a SW completely overlaps the area required by a user request, there is zero overlap error and we state that the SW has full coverage relevance to the request. Conversely, if the SW sensing area overlaps none of the coverage area of the request, the overlap error is 1, (the maximum error) and we say that the SW has zero coverage relevance to the query. For partial coverage, the error lies somewhere in the interval [0, 1].

As an example, we use the three SWs from Section 3.1 with coverage as shown in Table 2. We expand the user request into two separate requests: R_1 - Return the temperature observed at the location: Latitude 42.350, Longitude -83.020, and R_2 - Return the temperature observed by all sensors within the area bounded on the upper left by Latitude 42.500, Longitude -83.280 and Lower Right Latitude 42.350, Longitude -83.010. We want to know the relevance of the three SWs with respect to the two user requests.

Table 2: Coverage Boundaries for Observations of 3 SWs

SW ID	Upper Left Lat	Upper left Lon	Lower Right Lat	Lower right Lon
SW_1	42.531	-82.283	42.346	-83.030
SW_2	42.400	-83-080	42.341	-83.020
SW_3	42.475	-83.090	42.340	-83.030

Based on the above definitions and data, the coverage overlap errors and Coverage Relevance rankings for the three SWs with respect to R_1 and R_2 are shown in Table 3.

Table 3: Coverage Relevance of 3 SWs to requests R_1 & R_2

SW ID	R_1 Coverage Overlap Error	SW Rank with respect to R_1 coverage	R_2 Coverage Overlap Error	SW Rank with respect to R_2 coverage
SW_1	1.0	2	0.0740	1
SW_2	0.0	1	0.9136	3
SW_3	1,0	2	0.8148	2

3.3 Timing Metric: Measuring the Timing Relevance of a SW to an information request

The timing of a SW specifies the time or time period over which the SN can make observations and/or the frequency with which observations can be made. Given a SW and a user request for data about a specific time period, we want to be able to assess the degree to which the SW satisfies the timing requirements of the request. This is the timing relevance of the SW to the request. For real time queries, start times must be ≥ current time, and end times must be ≤ current time + anticipated remaining SW lifetime. For queries of archived data, start and end times are limited by historical date range of the data set.

We define the timing relevance metric in a manner similar to the coverage metric. If TR is the sensing time range required by the User Request, and TA is the portion of TR during which observations are actually made by the SW, then timing range overlap error, TOE_R is defined as:

$$TOE_R = 1.0 - TA/TR$$

We define the frequency relevance metric as follows. If FR is the observation frequency required by the user request, and if FA is the observation frequency of the SW, then the timing frequency overlap error is:

$$TOE_F = 1.0 - FA/FR$$

Continuing with the example from the previous section, suppose SW_1, SW_2, and SW_3 can make observations in accordance with the timing parameters presented in Table 4. Suppose, further that we have the following extension of user request, R_1: Return temperature observations every 10 minutes beginning at CT + 1000 hrs. and ending at CT + 2000 hrs. (where CT = current time).

Table 4: Timing Parameters for Observations of 3 SWs

SW ID	Observations per hour	Estimated Remaining Sensor Life (hrs)
SW_1	60	500
SW_2	12	3000
SW_3	2	5000

Based on the above definitions and data, the Timing Overlap error and Frequency Overlap error and Timing Relevance rankings for R are shown in Table 5.

Table 5: Timing and Frequency Relevance of 3 SWs to request R_1

SW ID	R_1 Timing Overlap Error	SW Rank with respect to R_1 Timing	R_1 Frequency Overlap Error	SW Rank with respect to R_1 Frequency
SW_1	1.0	2	0.0	1
SW_2	0.0	1	0.33	2
SW_3	0.0	1	0.50	3

3.4 Potential and Composition Relevance

As earlier noted, a SW may not have a high relevance factor as is, but may improve its relevance if modified. This could involve the tasking of nodes to move closer to an area of interest in order to improve coverage relevance. It might involve tasking of nodes to change their frequency of observation to improve timing relevance, or it could involve tasking of nodes to sense different phenomena that have shorter semantic distances to query terms in order to improve topical relevance. The potential of a SW to improve its relevance will be determined by examining the SW's capabilities metadata to determine how the SW's configuration or behavior can be modified to improve its relevance to an information request.

It is also possible that SWs could be combined so that the union of their nodes would provide more complete coverage and thus a higher coverage relevance than any of the component SWs. Possible relevance improvements would also be determined by examining SW metadata to assess composition options. The initial phase of this research will focus upon development of the relevance metrics defined in sections 3.1, 3.2 and 3.3. After some experience is gained in these areas, the research will be expanded to address potential and composition relevance.

4. AUTHORITY METRICS

Since SWs are not generally part of the surface web (i.e. SW data is not stored on web pages) the authority of a SW cannot be established by enumerating the number of links to other SW web pages of known authority. However, the sponsoring authority (the entity responsible for the creation and maintenance of the SW resource) is usually accessible on the surface web. In this project, we analyze the links to the sponsor to determine its authority with respect to the topic of an information request. We then assign that same authority to the SW resource (for that specific search topic). To determine the authority of the surface web accessible sponsoring authority, we use topic-sensitive PageRank, a mechanism that creates a bias rating for the most useful, accurate and authoritative sensor webs that relate to a set of representative topics. This is similar to the approach suggested by Taher Haveliwala in his extensive research on topic-sensitive PageRank [13] and is an extension of earlier research by Chakrabarti et al. [14] who demonstrated that web pages tend to point to other pages that relate to the same general topic. The topics in our case relate to sensor networks and sensor webs, and use an ontology that consists of knowledge, words, their meanings and their conceptual relationships to the most authoritative SW pages.

During an offline processing of the web crawl, we generate several topic-sensitive PageRank vectors; each biased using URLs from a top-level category of the Open Directory Project (ODP) [15]. At query time, the similarity of the query (and, if available, the query or user context) to each of these topics is calculated.

Let *Tj* be the set of URLs in the ODP category c_j. We then define the topic vector $\vec{p} = v_{ji}$ where:

$$v_{ji} = \begin{cases} \frac{1}{|T_j|} & i \in T_j, \\ 0 & i \notin T_j. \end{cases}$$

The PageRank vector for topic c_j will be $\vec{PR}(\alpha, \vec{v}_j)$ where the bias factor, α affects the degree to which the resultant PageRank vector is biased towards the topic vector $\vec{p}$..

The second step in the approach is performed at query time. Given a query *q*, let *q'* be the context of *q*. If the query was issued by highlighting the term *q* in some Web page *u*, then *q'* consists of the terms in *u*. For queries not done in context, let *q=q'*. Using a unigram language model, with parameters set to their maximum-likelihood estimates, we next compute the class probabilities for each of the 16 top level ODP classes, based on q'. Let q'_i be the *i*th term in the query (or query context) *q'*. Then given the query *q*, for each c_j we compute the following probability:

$$P(c_j|q') = \frac{P(c_j) \cdot P(q'|c_j)}{P(q')} \propto P(c_j) \cdot \prod_i P(q'_i|c_j)$$

A text index is then used to retrieve URLs for all documents containing the original query terms *q*. Next, we compute the query-sensitive importance score of each retrieved URL as follows. Let $rank_{jd}$ be the rank of document *d* identified by the rank vector $\vec{PR}(\alpha, \vec{v}_j)$ (i.e., the rank vector for topic c_j). For the Web document *d*, the query-sensitive importance score s_{qd} is defined as:

$$s_{qd} = \sum_j P(c_j|q') \cdot rank_{jd}$$

s_{qd} values will be normalized to the interval [0, 1] with a value of 1 indicating the maximum importance. Query results are ranked according to composite score s_{qd}. Following this approach will help avoid the problem of heavily linked pages receiving a high ranking for queries for which they have no particular topicality relevance or authority. This follows because pages considered important in some subject domains may not be considered important in others, regardless of what keywords may appear either in the page or in anchor text referring to the page.

5. PLANNED RESEARCH ACTIVITIES

In order to develop and refine the relevance and authority concepts presented in sections 3 and 4, a search model will be implemented that ranks a group of SWs with respect to a predefined set of information requests. We will identify the test set of deployed Sensor Networks by using SOS (Sensor Observation Service) resources such as National Oceanographic and Atmospheric Administration http://sdf.ndbc.noaa.gov/sos. We will then formulate a set of test queries that incorporate a variety of topical, coverage and timing criteria that might be satisfied by the test set of SWs.

A 4-dimension evaluation vector will be constructed for each of the candidate SWs with respect to each of the information requests. Elements 1-3 will consist of the 3-dimension relevance vector defined in section 3 (low values for the three vector components indicate high relevance). Element 4 will consist of the single dimension authority vector described in section 4 (high

value indicates high authority). For a given information request, the SW best able to satisfy the request has the lowest values in elements 1-3 and the highest value in element 4. A suitable scoring formula will be created to assign weights to evaluation vector components and calculate an overall SW ranking from these components. This task is probably best performed after initial gathering and examination of evaluation vector data.

Once a preliminary scoring formula is devised, a panel of test users will assess and evaluate the rankings of the SW resources returned by the model with respect to the various information requests. Model parameters, weights and the scoring methodology will be optimized based on feedback from the panel. In a future research phase, the issues of potential and composition relevance will be addressed and the evaluation vector will be expanded with an additional two elements to address these factors.

6. CONCLUSION

In this paper, we have compared and contrasted some of the key attributes of the surface web and the deep web, particularly as relates to obtaining data from wireless sensor networks and sensor webs. We have defined a three-dimension vector that provides metrics for quantifying the relevance of a sensor web to a specific query or request for data. We also defined an authority mechanism that does not just measure the overall popularity of a resource, but rather the interest in that resource with respect to a particular topic. For future research, we will be expanding the evaluation vector from four to six elements to incorporate potential and composition relevance. Ultimately, we will incorporate the above metrics into search engine dedicated to discovering and ranking of SW data resources.

7. REFERENCES

[1] Bergman, M. 2001. White Paper: the Deep Web: Surfacing Hidden Value. *The Journal of Electronic Publishing* vol. 7, no. 1.

[2] He, B., et.al. 2004. Accessing the Deep Web: A Survey. University of Illinois at Urbana.

[3] Wikipedia, Wireless Sensor Network: http://en.Wikipedia.org/wiki/Wireless_sensor_network.

[4] ON World Inc. 2009. Wireless Sensor Networks. http://www.onworld.com/.

[5] Wikipedia. Sensor Web, http://en.wikipedia.org/wiki/Sensor_Web

[6] OGC Sensor Web Enablement: Overview and High Level Architecture. Open Geospatial Consortium, Inc.

[7] OpenGIS Sensor Web Enablement Architecture. 2006. Open Geospatial Consortium, Inc.

[8] Resource description Framework (RDF). 2004. World Wide Web Consortium. www.w3.org/RDF.

[9] Barbosa, L., Freire, J. 2005. Searching for Hidden-Web Databases, *Eigth International Workshop on the Web and Databases (WebDb 2005)*. Baltimore, MD.

[10] Raghavan, S., Garcia-Molina, H. 2001. Crawling the Hidden Web. *Proceedings of the 27th VLDB Conference*. Rome, Italy.

[11] Corby, O. et al. 2005. Ontology-based Approximate Query Processing for Searching the Semantic Web with Corese. Institut National de Rescherche en Informatique et en Automatique

[12] Semantic Web for Earth and Environmental Terminology (SWEET). http://sweet.jpl.nasa.gov/ontology/

[13] Taher H. Haveliwala. 2003. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search., IEEE Vol. 15, N4. 4.

[14] Chakrabarti, S. et al. 2002. The structure of broad topics on the web. *Proceedings of the Eleventh International World Wide Web Conference*.

[15] The Open Directory Project, http://www.dmoz.org.