Intelligent Topic-Specific Web Crawling Using Sense-based Semantic Similarity Measures

Ali Pesaranghader [1], Norwati Mustapha [1], Ahmad Pesaranghader [2]

[1] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

[2] Faculty of Creative Multimedia, Multimedia University, Malaysia

ali.pgh@sfmd.ir, norwati@upm.edu.my, ahmad.pgh@sfmd.ir

***Abstract* — With the rapid growth of the Internet, looking for desired information on the Web has become a tedious and time consuming task. Topic-specific web crawlers, as automated web robots, address this problem by traversing the Web for collecting information related to our desired topics. Various methods have been proposed to develop these web robots. Nonetheless, they hardly take into account the desired sense of the given topic which certainly plays an essential role to find more relevant web pages. In this paper, we aim to improve topic-specific web crawling by disambiguating the desired sense of the input topic. This avoids crawling irrelevant links interlaced with other senses of the topic. For this purpose, by considering semantic of hypertexts, we utilize Lin semantic similarity measure in our crawler, called LinCrawler, to distinguish topic sense-related links from the others. Finally, we compare LinCrawler with TFCrawler which only considers terms frequencies in the hypertexts. Experimental results show LinCrawler surpasses TFCrawler in collecting more relevant web pages.**

Keywords – Topic-Specific Web Crawling; Link Prediction; Information Retrieval; Web Data Mining; Semantic Web.

I. INTRODUCTION

Today, the Internet is the main door to the world of information. People all over the world are using public searching tools such as general search engines to find their desired web pages. Typically, general search engines are not able to answer users' requests accurately and only top ranked links are considered valuable by web surfers. On the contrary, web directories, also known as directory search engines, provide more relevant information by categorizing web pages into different directories. Nevertheless, building these directories manually is an impractical and tedious task. Therefore, having automated tools able to identify web pages and categorize them is essential. In this regard, topic-specific crawlers (also known as focused crawlers in some applications), as the hearts of vertical search engines, have been devised to traverse the Internet and retrieve relevant web pages with high precision by constraining the scope of the crawl. For this purpose, topic-specific crawlers analyze the contents of web pages, structures of links among web pages, or combination of both of them. Fig. 1 is an example that shows the topic-specific web crawler discovers and collects more relevant web pages compared to the Breadth-First (BF) crawler. In this figure, thick edges are traversed by the crawlers in the order of numbers on them.

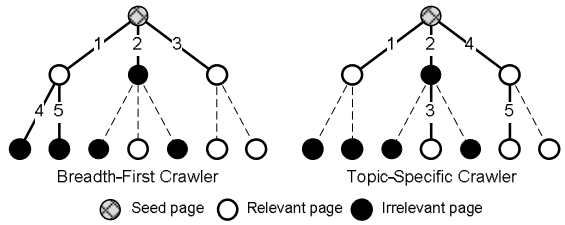

Fig 1. Topic-Specific Crawler vs. Breadth-First Crawler

Most of the times, the given topic carries more than one meaning which cause misleading the crawler to collect irrelevant web pages [1]. Thus, empowering topic-specific crawlers with techniques that make them be aware of senses of topics is vital. In this paper, we propose a crawler equipped with a semantic similarity measure that makes it capable of taking senses of input topics into account.

The remainder of this paper is organized as follows: Section II reviews related works. We explain textual and semantic similarity measures which are practical for topic-specific web crawling in Section III. Section IV describes our contribution and how a sense-based crawler works. We introduce our framework in Section V. In Section VI, we conduct our experiments to compare textual and semantic topic-specific crawlers. Finally, we conclude in Section VII.

II. RELATED WORK

We categorize topic-specific and focused crawlers into four classes: 1) content-based, 2) link-based, 3) content and link-based, and 4) ontology-based crawlers.

Crawlers based on Content Analysis – These crawlers employ machine learning techniques to make decisions about relevance of web pages to a given topic. Pant et al. [2] showed that Support Vector Machine (SVM) and Neural Networks (NN) are notable choices compared to the Naive Bayes. Li et al. [3] claimed Subspace Methods and Naive Bayes outperform Nearest Neighbor and Decision Tree to classify seven groups of Yahoo! News. McCallum et al. [4] designed an intelligent crawler based on reinforcement learning and text classification for collecting scientific papers on the Web. This group of topic-specific crawlers is highly domain dependent as training effective classifiers is crucial.

Crawlers based on Link Analysis – These crawlers focus on the popularity of web pages rather than their contents. Here, popularity means how many times a web page is

referenced by the other web pages. Generally, a web page pointed many times by other web pages is more important than a remote web page. This notion is known as BackLink method. PageRank [5 and 6], effectively used in Google search engine, is an intelligent method that takes into account not only the number of pointing links to a page but also how important they are. In Hyperlink-Induced Topic Search (HITS) method by Kleinberg [7], each web page has two scores: *authority* and *hub*. It works based on this idea that a good authority should be referenced by a number of good hubs, and a good hub should refer to a number of good authorities. Eventually, the web pages with high authority scores are selected as relevant web pages. Feng et al. [8] introduced the Navigational Rank (NR) method which has a significant performance when the seeds are far away from the target pages. Even though, this group is independent of domain, but it is not reactive in time to the continuous changes in the structure of the Web.

Crawlers based on Content and Link Analysis – These crawlers find potential links pointing to relevant web pages by perusing the identified relevant web pages. Usually, they are not only domain independent but also invulnerable to the dynamic structure of the Web. Almpanidis et al. [9] introduced the Hypertext Combined Latent Analysis (HCLA) for the content and link analysis. They indicated the crawler based on HCLA outperforms BackLink, Breadth-First, PageRank and two extensions of Shark-Search [10] crawlers. However, it requires more processing power and memory resources. Chen et al. [11] introduced the HAWK focused crawler. First, HAWK checks whether the web page crawled is relevant. Then, in the relevant case, it uses Shark-Search algorithm to predict relevant outgoing links. Yin et al. [12] designed the BBF crawler which uses general search engine and χ^2 to find informative terms related to the input type. Then, it discovers potential relevant links by applying a C4.5 decision tree. Their results showed the superiority of BBF over Best-First and Breadth-First crawler. Pesaranghader et al. [13 and 14] proposed Term Frequency-Information Content (TF-IC) weighting measure to improve multi-term topic focused crawling. They showed that the TF-IC crawler outperforms two crawlers based on Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Semantic Indexing (LSI).

Crawlers based on Ontologies – These crawlers avoid crawling irrelevant web pages to a given topic by calculating semantic similarity between hypertexts' terms and the topic. LSCrawler [15] by Yuvarani, to find relevant links, measures the similarity between terms around links (hypertexts) and a given topic through calculating the distances of their corresponding entities in an ontology. In LSCrawler, an ontology repository keeps different ontologies and each of which might be compatible with the given topic. Evaluation of LSCrawler lacks detail information which cannot show its effectiveness clearly. Ganesh [16] suggested using categories of web directories to build a hierarchy of entities (or ontology). Crawlers can use this hierarchy to find a knowledge-path from the root to the corresponding category of the input topic. Finally, this knowledge-path is used to collect web pages from the root (the most general category) to corresponding category of the given topic (the most specific category). This method left without any evaluation. Lokman [17] identifies relevant links by considering entities in the Unified Medical Language System (UMLS) [1] metathesaurus. First, Lokman finds all associated entities with the input topic and then uses them to score links in the web pages. Associated entities are in four groups of exact terms, synonyms, partial relevance, and contextual relevance. Also, it considers the frequencies of ontological entities in a web page to assign a score to that page. Finally, scores of links and web pages are used to prioritize links before adding them into the frontier. They showed that Lokman outperforms Best-First crawler. Even though, their method seems promising but they unclearly mentioned how they find associated entities related to an input topic. Topic-specific crawlers based on ontologies are able to find more relevant web pages by considering ontological entities and other similar terms related to the input topic. In addition, they can be combined with other three groups. However, current methods are not able to consider meanings of input topics.

In overall, three first groups suffer from lack of ability in considering semantic similarity amongst terms and the topic. On the other hand, the forth group takes synonyms and other related terms (or entities) to the topic into account using ontologies. Nonetheless, the need for having integrated semantic similarity measures is still felt.

Along with considering semantic relationships, semantic crawlers need to consider the sense of the input topic as well even if the applied ontology is a domain-specific one. For instance, as shown in Table I, the term "auricle" carries two senses (or concepts). Without specifying the sense of topic, the crawler is exposed to archive web pages which are related to the other sense. Thus, providing a method that considers the desired sense of input topics is crucial.

TABLE I. SENSES OF THE TERM "AURICLE"

Sense	Description*
1	***A small conical pouch projecting from the upper anterior part of each atrium of the heart.***
2	***The externally visible cartilaginous structure of the external ear.***

*Descriptions are brought from WordNet.

In this paper, we propose a crawler which not only considers the semantic similarity between hypertexts' terms and the input topic but also the desired sense of the input topic. Before showing the dignity of our crawler over textual crawlers, we describe textual (lexical) and semantic similarity measures separately in the next section.

III. SIMILARTY MEASURES

We categorize similarity measures into two groups: 1) textual (or lexical) similarity measures, and 2) semantic similarity measures. Both are widely used in Information Retrieval (IR) and Natural Language Processing (NLP).

[1] http://www.nlm.nih.gov/research/umls/

A. Textual Similarity Measures

Textual (lexical) similarity measures are used to calculate the similarity between two documents. For this purpose, documents are represented as vectors in the space of terms, i.e. a term-document matrix. This way of representation is also known as Vector Space Model (VSM) in IR. For topic-specific crawling, the input topic and each web page, or some parts of it, are considered as documents to be shown as vectors.

Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) are two common weighting measures to assign a weight to each of terms that occur in a document. Since TF-IDF is often used to prioritize specific terms out of the other terms in multi-term topics crawling [14], in this paper, we only focus on absolute frequency of terms and apply TF measure to represent documents as vectors. A document vector is represented by (1), where $tf_{i,j}$ is frequency of the term i in the document j. In topic-specific crawling based on content and link analysis, the text around links (hypertexts) are considered as short documents. Also, the topic vector is shown as (2).

$$\vec{d}_j = (tf_{1,j}, tf_{2,j}, ..., tf_{n,j}) \quad (1)$$

$$\vec{q} = (tf_{1,q}, tf_{2,q}, ..., tf_{n,q}) \quad (2)$$

After representation of the documents and the input topic as vectors, it is possible to calculate the similarity between a document (here a hypertext) and the input topic by applying cosine similarity measure (3). It computes the similarity between a document vector and the topic vector by considering magnitude of the angle between them. Smaller angle means greater similarity.

$$sim(\vec{d}, \vec{q}) = \cos\theta = \frac{\vec{d} \cdot \vec{q}}{\left|\vec{d}\right| \times \left|\vec{q}\right|} \quad (3)$$

Textual similarity measures do not consider semantic relationships between terms and the input topic which causes losing a considerable amount of information. In the next subsection, we describe semantic similarity measures as remedies to solve this issue.

B. Semantic Similarity Measures

Semantic similarity measures, as a subset of techniques in Computational Linguistics, are used to create groups of similar terms automatically by utilizing existing ontologies and/or big corpora. The output of a similarity measure is a value showing how much two given terms are semantically similar. These measures are successfully applied in a wide range of applications [18 and 19]. In [20], we categorize semantic similarity measures into three groups: 1) Similarity measures based on Path, 2) Similarity measures based on Path and Depth, and 3) Similarity measures based on Path and Information Content. In addition to semantic similarity measures, based on the application, semantic relatedness measures also can be applied for developing topic-specific web crawlers. Generally, in theory, the related terms for a specific term are those accompany that term in a corpus. These related terms could be considered as representatives of the topic for crawling. In this paper, we only focus on semantic similarity measures.

Similarity Measures based on Path – These measures use distance of concepts from each other in a taxonomy to calculate their similarities. In other words, the length of shortest path from two concepts indicates how similar they are. Caviedes [21] and Rada [22] proposed their semantic similarity measures based on the idea of shortest path among concepts. In a taxonomy, concepts in upper levels are abstract in their meaning and so less similar with others. On the other hand, concepts in the lower levels are concrete in their meanings and more similar especially when they are siblings. The possibility of high similarity among upper concepts, because of small number of jumps, is drawbacks of these path-based measures.

Similarity Measures based on Path and Depth – These measures are proposed to overcome the drawback of first group. In this regard, they use path and depth of concepts to calculate their similarities. Wu and Palmer [23], Leacock [24] and Zhong [25] measures are examples of this group. Nevertheless, frequency of concepts in corpora is another notable factor for measuring the similarity of two concepts. When one concept is less frequent than the other concepts, it should be considered more informative and concrete in its meaning. This group suffers from overlooking frequencies of concepts which leads to the usage of the third group.

Similarity Measures based on Path and Information Content – As we mentioned above, similarity measures in the second group are weak as they do not consider concepts' frequencies in their calculation of similarity between two concepts. In overall, low frequent concepts are much more informative and concrete in their meanings than high frequent concepts. This notion is known as Information Content (IC) in Information Theory society. Resnik [26], Jiang [27] and Lin [28] measures use concepts' information contents to compute the similarity among them.

Generally, similarity measures in the third group yield better results in comparison with the first and second groups. In this paper, our focus is on the third one.

After becoming familiar with the textual and semantic similarity measures; in the next section, we describe how semantic similarity measures are applied into topic-specific web crawling for obtaining more relevant web pages.

IV. Using Semantic Similarity Measures to Develop Intelligent Topic-Specific Web Crawlers

Semantic Similarity measures can be applied to improve topic-specific crawling by calculating semantic similarity between hypertexts and the input topic with respect to the

desired sense of the topic. In this paper, we are going to exploit Lin semantic similarity measure for this purpose. The similarity between two concepts using Lin measure is calculated by (4).

$$Sim_{Lin}(C_1, C_2) = \frac{2 \times IC(LCS(C_1, C_2))}{IC(C_1) + IC(C_2)} \quad (4)$$

Where $IC(C_1)$ and $IC(C_2)$ are information contents of C_1 and C_2 respectively. $LCS(C_1, C_2)$ is the least common subsumer of two concepts and $IC(LCS(C_1, C_2))$ is its information content. The information content of concept C_i is calculated by (5).

$$IC(C_i) = -\log\left(\frac{CF(C_i) + IF(C_i)}{N}\right) \quad (5)$$

Where $CF(C_i)$ is frequency of the concept C_i in the corpus and $IF(C_i)$ is its inherited concept frequency, i.e. sum of its descendants' frequencies, and N is sum of the augmented frequencies of all concepts (frequencies of concepts and their descendants) in the taxonomy.

For instance, consider the input topic is "*iris*" and we are about to collect web pages which are related to the medical sense of *iris* (a part of an eye not a flower). Hereafter, we show a topic with specific sense as a <*topic:sense*> pair. For this example, it is <*iris:medical*>. Thus, the task of the crawler equipped with the Lin measure is predicting links related to the <*iris:medical*>. Table II shows the similarity scores of terms occurred in a hypertext with the specified sense of input topic. For example, the similarity score between *plant* and <*iris:medical*> is 0.08. As this table shows, the semantic crawler avoids link 1 related to the gardening while it finds link 2 is more relevant to the medical sense of the topic. In other words, a link is predicted relevant if its corresponding hypertext's score calculated with the desired sense is greater than the other scores calculated with other senses.

TABLE II. AN EXAMPLE SHOWING HOW SEMANTIC CRAWLER TAKES INTO ACCOUNT THE TOPIC'S SENSES

Link	Hypertext Keywords' Scores		
1	**[plant, leaf, flower]**	**Topic Sense**	**Score***
	[0.15, 0.14, 0.53]	gardening	0.273
	[0.08, 0.07, 0.08]	**medical**	0.077
2	**[eye, sphincter]**	**Topic Sense**	**Score**
	[0.09, 0.07]	gardening	0.08
	[0.41, 0.36]	**medical**	**0.385**

* Here, total score is equal to average of all scores.

In the next section, we describe our Web Sapphire Collector II (WSC II) framework which comprises different crawlers. In this paper, we only focus on three of them: BFCrawler, TFCrawler and LinCrawler. TFCrawler is a textual crawler which only looks at terms lexically while LinCrawler also looks at them semantically.

V. WEB SAPPHIRE COLLECTOR II FRAMEWORK

In order to accomplish this research, we implemented our own topic-specific crawling framework. We called it Web Sapphire Collector II. The outline of this framework is depicted in Fig. 2. Generally, Frontier, Fetcher and Extractor are the most important components of a crawler. The Frontier holds the URLs to be crawled in a near future. The Fetcher downloads and fetches web pages from the Internet. Extractor peruses the web pages downloaded and elicits potential links. Then, candidate URLs, i.e. potential links discovered by Extractor, are sent to the frontier queue. As Fig. 2 represents there are five common processes in all crawlers while their LFCs' components work differently. Those five steps are as follows:

1. Getting the input topic from user via an UI and then start to collect *k* seed pages from general search engines as Google using the provided APIs.
2. Adding seed pages into the Frontier queue.
3. Receiving new URLs from Frontier and fetching them from the Internet.
4. Extracting URLs from web pages by overlooking the repeated ones. Then, sending the new URLs and their hypertext (in a predefined window size) to the Link Filtering Component (LFC) which works differently based on the type of the crawler.
5. Archiving web pages downloaded.

After these five steps, the LFC filters irrelevant links out and sends the new top scored links to the frontier queue. Then, the progress above is reiterated until the crawler meets a predefined constraint, e.g. total number of pages to be crawled.

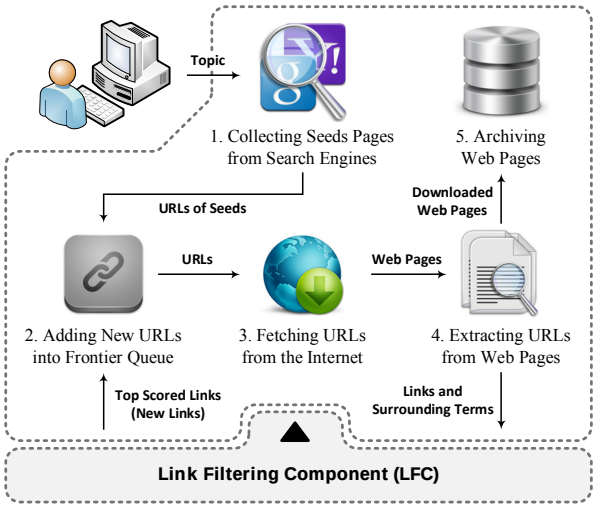

Fig 2. Overview of WSC II

In this paper, we have concentrated on three crawlers called BFCrawler (Breadth-First Crawler), TFCrawler and LinCrawler. Expect BFCrawler, TFCrawler and LinCrawler have their own specific Link Filtering Component (LFC) to score the links, filter out irrelevant ones, and then sort the relevant ones before adding them into the frontier queue.

BFCrawler – This crawler is the simplest crawler in our framework without any special computational technique. The crawler directly adds the new URLs into the frontier queue after extracting them from the web pages. In other words, there is no LFC in this crawler.

TFCrawler – This crawler scores the links based on their hypertexts' textual similarities with the input topic. Links' hypertexts, in a predefined window size, are considered as short documents to be represented as vectors (1). Fig. 3 represents the steps that TFCrawler adopts by its LFC:

1. (Step 6) Building the vectors of links after removing stop word and stemming terms.
2. (Step 7) Getting cosine similarity measure between each of links' vectors and the topic's vector.

Then the outputs of step 7 are the top scored links which will be sent to the frontier queue.

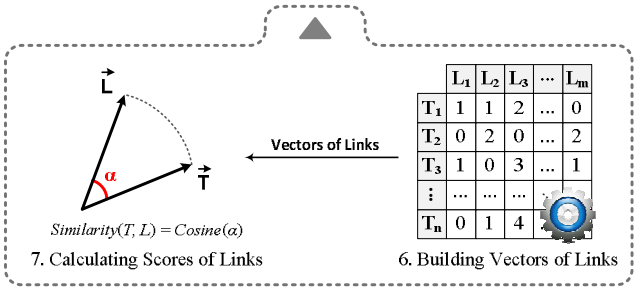

Fig 3. Inside LFC of TFCrawler

LinCrawler – This crawler scores the links based on semantic similarities of their hypertexts and the input topic through considering its senses. Fig. 4 shows the steps that LinCrawler adopts by its LFC:

1. (Step 6) Building sets of terms for each link from their corresponding windows after removing the stop words and finding terms roots.
2. (Step 7) Calculating the augmented similarity of terms' sets and the topic for each of its senses. If the score of a link calculated with desired sense is the greatest one, it will be considered as a relevant link.

Then the outputs of step 7 are the top scored links which will be sent to the frontier queue.

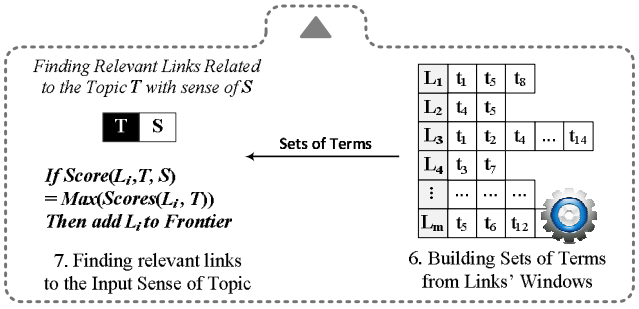

Fig 4. Inside LFC of LinCrawler

The following pseudo code also shows how the scores of links are computed by LinCrawler.

```
getLinksScores (linksTermsSets, topic, sense) {
  double[] scores;
  for (i : 0 to i : linksTermsSets.length - 1)
    scores[i] = getLinSim(linksTermsSets[i], topic, sense);
  return scores;
}
getLinSim (termsSet, topic, sense){
  double[] sensesScores;
  int numSenses = getTopicNumSenses(topic);
  for (i : 0 to i : numSenses - 1){
    int k = termsSet.length;
    for ( j : 0 to j : k - 1)
      sensesScores[i] += getLin(topic, i + 1, termsSet[j]) / k;
  }
  int score = 0;
  int maxScoreIndex = getMaxScoreIndex(sensesScores);
  if (sense = maxScoreIndex + 1)
    score = sensesScores[sense - 1];
  return score;
}
```

Code I. Pseudo Code for Calculating Links' Semantic Similarity Scores

VI. Experiments and Discussions

We conduct our experiments on three topics having different senses from the computer, medical and sport domains. They are listed in Table III. For each topic's sense, we run each of BFCrawler, TFCrawler and LinCrawler separately to collect 6000 web pages, and then we compare their performance in terms of their harvest ratios results.

Table III. Topics Chosen for Experiments

Topic	Sense	Brief Description*
Cookie <*Confectionery*>	1	***Any of various small flat sweet cakes.***
Cookie <*Internet*>	2**	***A short line of text that a web site puts on your computer's hard drive when you access the web site.***
Iris <*Gardening*>	1	***Plants with sword-shaped leaves and erect stalks bearing bright-colored flowers composed of three petals and three drooping sepals.***
Iris <*Medical*>	2	***Muscular diaphragm that controls the size of the pupil.***
Squash <*Vegetal*>	1, 2	***Sense 1: Any of numerous annual trailing plants of the genus cucurbita grown for their fleshy edible fruits.***
		Sense 2: Edible fruit of a squash plant; eaten as a vegetable.
Squash <*Sport*>	3	***A game played in an enclosed court by two or four players who strike the ball with long-handled rackets.***

*Descriptions are brought from WordNet.

** Sense 3 in WordNet.

A. *Experiment Taxonomy and Data*

We chose WordNet[2] as the taxonomy which comprises concepts along with relationships among them. Meanwhile, we used pre-computed information contents from the SemCor corpus[3]. These ICs have been calculated based on the relationships in the WordNet.

B. *Evaluation*

In order to measure the similarity or relatedness of the web pages crawled we create a file called Oracle. The Oracle file contains definition of topic from Wikipedia and WordNet as well as definitions of its children in WordNet. Meanwhile, the topic term is eliminated in Oracle file to check whether crawlers collected web pages related to the correct sense or not. Then, as Fig. 5 illustrates, we measure the similarity between Oracle file and web pages crawled. The web pages with similarity score of equal or greater than a predefined minimum score δ, e.g. $\delta=0.2$, are considered relevant. Eventually, we could compare performances of the crawlers after computing scores of web pages crawled.

To compare the performance of crawlers, Harvest Ratio (HR) measure is employed (6). As we see in the equation, it is the total number of relevant web pages crawled over the total number of web pages crawled.

$$HR = \frac{\| relevant\ web\ pages\ crawled \|}{\| total\ web\ pages\ crawled \|} \tag{6}$$

HR shows how effective a crawler is to collect relevant pages. Having a high HR is always preferred.

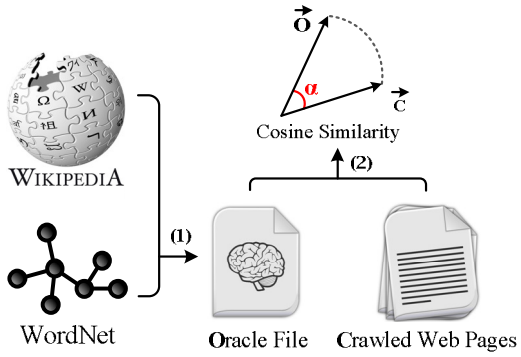

Fig 5. Overview of Evaluation Process Method

Moreover, in order to show significance of our crawler, we use statistical paired samples t-test to calculate p-values for each pair of crawlers. As stated in (7) and (8), H_0 is the null hypothesis and H_1 is the alternative one. $D(\varepsilon,\beta)$ is the difference between harvest ratios of crawlers ε and β, and $\mu_{D(\varepsilon,\beta)}$ is the average of differences. Meanwhile, μ_0 is the expected difference average and it is equal to zero in our experiments.

[2] http://wordnet.princeton.edu/
[3] http://wn-similarity.sourceforge.net/

$$H_0 : \mu_{D(\varepsilon,\beta)} \le \mu_0 \tag{7}$$

$$H_1 : \mu_{D(\varepsilon,\beta)} > \mu_0 \tag{8}$$

The value of t-test for paired samples is calculated by (9). $S_{D(\varepsilon,\beta)}$ is the standard deviation of the differences, and n is the number of samples.

$$t = \frac{\mu_{D(\varepsilon,\beta)} - \mu_0}{S_{D(\varepsilon,\beta)}} \times \sqrt{n} \tag{9}$$

The $\mu_{D(\varepsilon,\beta)}$ and $S_{D(\varepsilon,\beta)}$ are calculated by (10) and (11) respectively.

$$\mu_{D(\varepsilon,\beta)} = \frac{\sum_{i=1}^{n} D_i(\varepsilon,\beta)}{n} \tag{10}$$

$$S_{D(\varepsilon,\beta)} = \sqrt{\frac{1}{n-1} \times \sum_{i=1}^{n} (D_i(\varepsilon,\beta) - \mu_{D(\varepsilon,\beta)})^2} \tag{11}$$

After calculating value of t for right-tailed test, we can compute the *p-value* using t and *degree of freedom* ($df = n-1$) for crawlers ε and β. T-Distribution Table could be used to estimate the p-value as well. If the p-value is greater than the significance level (α) or value of t is less than t_α, we fail to reject null hypothesis which means there is no strong evidence to show crawler ε works better than crawler β. Otherwise, we reject null hypothesis which means crawler ε outperforms crawler β. Typically, 0.05 is considered for value of α. Fig. 6 shows rejection area for the null hypothesis.

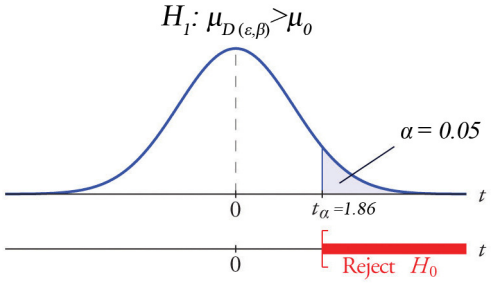

Fig 6. Rejection Area for Null Hypothesis in Right-tailed T-test

C. *Experiment on Internet×Confectionary Topic*

We conduct our first experiment on a topic from the Internet world which is "*Cookie*". As Table III listed, in the Internet world, *cookie* is a piece of information stored in the cache of users' internet browsers when users visit websites. On the general side, *cookie* is any kind of small flat sweet biscuit or cake. We run LinCrawler twice with the different senses of this topic. Then, for each sense, we compare its performance against BFCrawler and TFCrawler. The results of experiments are shown in Tables IV and V.

TABLE IV. RESULT OF EXPERIMENT ON <COOKIE:INTERNET> PAIR

	LinCrawler	TFCrawler	BFCrawler
HR	**0.351**	0.179	0.118

As Table IV shows, LinCrawler outperforms two other crawlers for collecting more relevant web pages to the <*Cookie:Internet*> pair. The performance of LinCrawler is 23% and 17% greater than that of BFCrawler and TFCrawler respectively. Fig. 7 represents the trend of each crawler's harvest ratio per percentage of pages crawled.

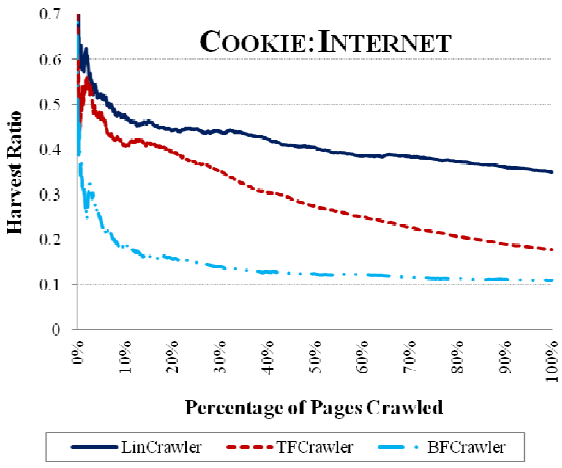

Fig 7. Comparing HRs of Crawlers for <Cookie:Internet> Pair

Table V represents the HR of each crawler for the <*Cookie:Confectionary*> pair.

TABLE V. RESULT OF EXPERIMENT ON <COOKIE:CONFECTIONARY> PAIR

	LinCrawler	TFCrawler	BFCrawler
HR	**0.618**	0.147	0.034

The trend of each crawler's harvest ratio is represented in Fig. 8. This figure shows clearly the dignity of LinCrawler over TFCrawler and BFCrawler.

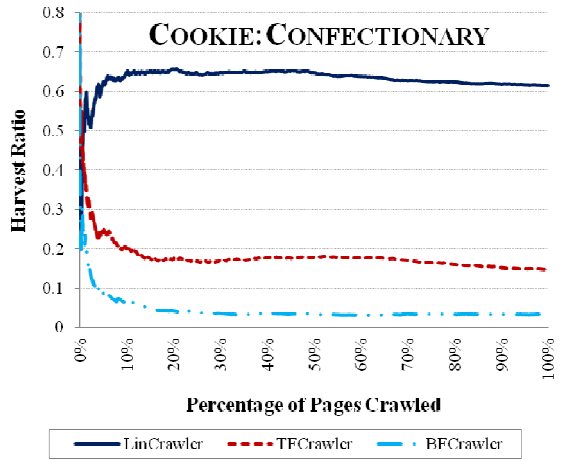

Fig 8. Comparing HRs of Crawlers for <Cookie:Confectionary> Pair

D. *Experiment on Medical×Gardening Topic*

To conduct our second experiment, we have chosen another disambiguate topic term from the medical world which is "*Iris*". Descriptions of two common senses of this topic are brought in Table III. In the medical domain, *iris* is a muscular diaphragm that controls the size of pupil; while in the gardening domain, *iris* is a kind of flower. As our first experiment, we run LinCrawler for each sense of the *iris* topic. Tables VI and VII compare the performances of crawlers for the *medical* and *gardening* senses of this topic.

TABLE VI. RESULT OF EXPERIMENT ON <IRIS:MEDICAL> PAIR

	LinCrawler	TFCrawler	BFCrawler
HR	**0.253**	0.031	0.058

LinCrawler outperforms BFCrawler and TFCrawler for the <*Iris:Medical*> pair. Performance of LinCrawler is eight and four times greater than that of TFCrawler and BFCrawler. We can see dominance of the LinCrawler over two other crawlers in Fig. 9. The interesting point is better performance of the BFCrawler compared to the TFCrawler which indicates less publicity of *iris* with the sense of medical on the Web.

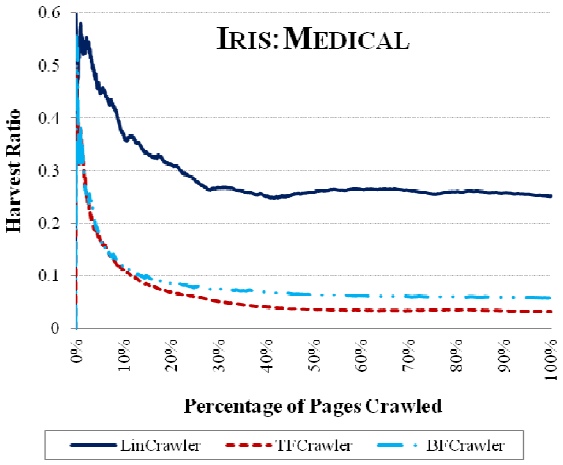

Fig 9. Comparing HRs of Crawlers for <Iris:Medical> Pair

Table VII represents LinCrawler's performance is 17% greater than TFCrawler's for collecting web pages related to the <*Iris:Gardening*> pair. Fig. 10 represents the trends of crawlers' HRs. In this figure, the superiority of LinCrawler is clearly seen.

TABLE VII. RESULT OF EXPERIMENT ON <IRIS:GARDENING> PAIR

	LinCrawler	TFCrawler	BFCrawler
HR	**0.459**	0.288	0.059

Comparing TFCrawler's performances for two senses of *iris* implies the more publicity of the first sense (a kind of flower) rather than the second one (a part of an eye) on the Internet.

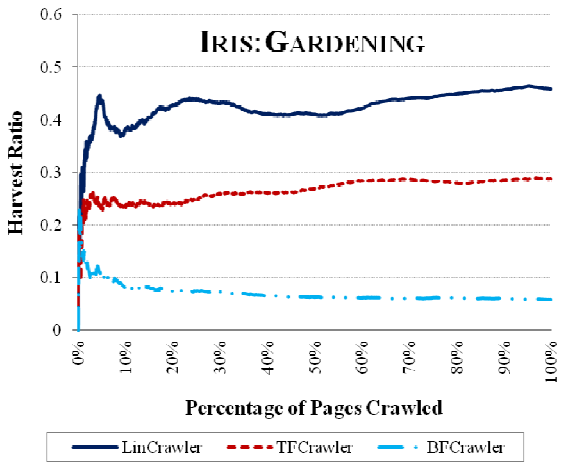

Fig 10. Comparing HRs of Crawlers for <Iris:Gardening> Pair

E. *Experiment on Sport×Vegetal Topic*

For further evaluation of our crawler, we have picked another topic from the sport world. The topic is "*Squash*" which has three senses. One sense is related to the sport, and the others are related to the vegetable world. A brief description for each sense has been stated in Table III. Tables VIII and IX compare the harvest ratios of crawlers for the <*Squash:Vegetal*> and <*Squash:Sport*> pairs. For each pair, the LinCrawler outperforms two other crawlers. In Table VIII, we can see the performance of LinCrawler is 33% and 36% greater than the performances of TFCrawler and BFCrawler respectively.

TABLE VIII. RESULT OF EXPERIMENT ON <SQUASH:VEGETAL> PAIR

	LinCrawler	TFCrawler	BFCrawler
HR	**0.437**	0.105	0.072

The trend of each crawler's harvest ratio for the <*Squash:Vegetal*> pair is presented in Fig. 11 and the superiority of LinCrawler over two other crawlers is clearly seen.

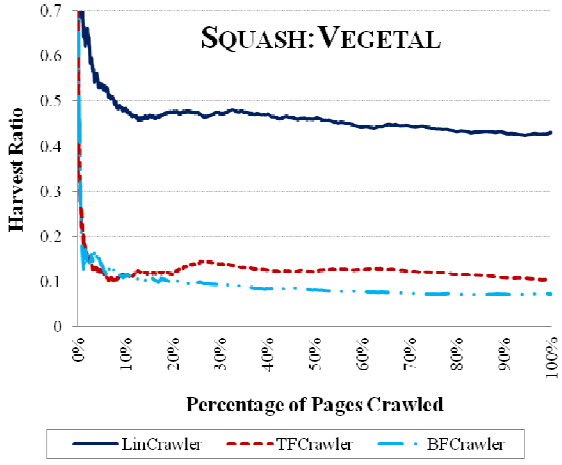

Fig 11. Comparing HRs of Crawlers for <Squash:Vegetal> Pair

The superiority of LinCrawler over TFCrawler and BFCrawler is obvious for the <*Squash:Sport*> pair. Its harvest ratio is three times greater than TFCrawler's.

TABLE IX. RESULT OF EXPERIMENT ON <SQUASH:SPORT> PAIR

	LinCrawler	TFCrawler	BFCrawler
HR	**0.56**	0.181	0.115

Fig. 12 represents trends of harvest ratios for three crawlers per percentage of web pages crawled for the <*Squash:Sport*> pair.

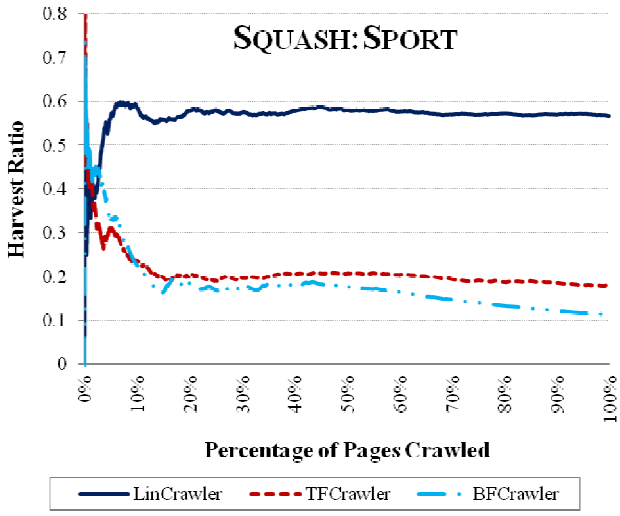

Fig 12. Comparing HRs of Crawlers for <Squash:Sport> Pair

F. *Statistical Hypothesis Testing*

Having harvest ratios of crawlers for six <topic:sense> pairs, presented in Table X, makes us able to do paired sample t-test to see whether LinCrawler has a better performance compared to the TFCrawler and BFCrawler.

TABLE X. HARVEST RATIOS OF CRAWLERS FOR EACH <TOPIC:SENSE> PAIR

	$HR_{LinCrawler}$	$HR_{TFCrawler}$	$HR_{BFCrawler}$
<Cookie:Confectionary>	0.618	0.147	0.034
<Cookie:Internet>	0.351	0.179	0.118
<Iris:Gardening>	0.459	0.288	0.059
<Iris:Medical>	0.253	0.031	0.058
<Squash:Vegetal>	0.437	0.105	0.072
<Squash:Sport>	0.56	0.181	0.115

Table XI represents the differences between harvest ratios of each paired crawlers to be compared with each other. For instance, the difference between harvest ratio of LinCrawler and TFCrawler for the <Iris:Medical> pair is 0.222. Average (μ_D) and standard deviations (S_D) of differences have been calculated for each pair (each column) and the results are available in this table.

TABLE XI. DIFFERENCES BETWEEN HARVEST RATIOS OF CRAWLERS FOR EACH <TOPIC:SENSE> PAIR

	$D(\varepsilon,\beta)$		
	***D*(*Lin*, *BF*)**	***D*(*Lin*, *TF*)**	***D*(*TF*, *BF*)**
<Cookie:Confectionary>	0.584	0.471	0.113
<Cookie:Internet>	0.233	0.172	0.061
<Iris:Gardening>	0.4	0.171	0.229
<Iris:Medical>	0.195	0.222	-0.027
<Squash:Vegetal>	0.365	0.332	0.033
<Squash:Sport>	0.445	0.379	0.066
Average (μ_D)	0.37	0.291	0.079
Standard Deviation (S_D)	0.143	0.122	0.087

By placing values of μ_D and S_D for each paired crawlers in (9), we can calculate value of t-test for that pair. The values of t-tests have been listed in Table XII.

TABLE XII. VALUES OF T-TESTS

	BFCrawler	**TFCrawler**
LinCrawler	5.192343	4.750084
TFCrawler	1.828256	–

By considering values of t-tests, having right-tailed tests and degrees of freedom, we can compute p-values for each paired crawlers. Table XIII contains p-values for each two crawlers compared with each other.

TABLE XIII. P-VALUES OF PAIRED CRAWLERS

	BFCrawler	**TFCrawler**
LinCrawler	0.001745	0.002552
TFCrawler	0.063526	–

The p-values calculated for {LinCrawler, BFCrawler} and {LinCrawler, TFCrawler} pairs are considerably less than significance level (α) which means we reject null hypothesis and there is strong evidence to accept the LinCrawler has a dominant performance over two other crawlers. On the other hand, the p-value calculated for {TFCrawler, BFCrawler} pair is greater than significance level meaning we failed to reject null hypothesis since there is no strong evidence to show TFCrawler works better than BFCrawler.

VII. CONCLUSION AND FUTURE WORK

Textual topic-specific web crawlers overlook huge amount of information related to the topics because of lack of ability in considering semantic similarities and relatedness among terms in web pages. Besides, senses of given topic play an important role to collect related web pages correctly. In this regard, in this paper, we proposed using semantic similarity measures along with specifying desired sense of input topic to empower topic-specific crawlers. We implemented our semantic topic-specific crawler, named LinCrawler, by applying Lin semantic similarity measure and using concepts' relationships in the WordNet. Experimental results proved the superiority of LinCrawler over TFCrawler (the crawler works lexically) and BFCrawler (the Breadth-First crawler).

For future work, implementing and comparing other semantic topic-specific crawlers based on depth-based semantic similarity measures [23-25], and other information content-based semantic similarity measures [26 and 27] is possible. In addition, comparing ontological crawler, as LinCrawler, with BBF crawler [12] is worth to conduct.

REFERENCES

[1] A. Pesaranghader, N. Mustapha, and A. Pesaranghader, "Applying Semantic Similarity Measures to Enhance Topic-Specific Web Crawling", Proc. 13th International Conference on Intelligent Systems Design and Applications, Dec. 2013, pp. 205-212.

[2] G. Pant and P. Srinivasan, "Learning to Crawl: Comparing Classification Schemes", ACM Transaction on Information System, vol. 23, Dec. 2005, pp. 430-462, doi: 10.1145/1095872.1095875.

[3] Y. H. Li and A. K. Jain, "Classification of Text Documents", The Computer Journal, Oxford University Press, vol. 41, no .8, 1998, pp. 537-546, doi: 10.1093/comjnl/41.8.537.

[4] A. Mccallum, K. Nigam, J. Rennie, and K. Seymore, "A Machine Learning Approach to Building Domain-Specific Search Engines", Proc. International Joint Conference on Artificial Intelligence (IJCAI'99), Morgan Kaufmann Pub., vol. 2, Aug. 1999, pp. 662-667.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bring Order to the Web", Technical Report, Stanford University, 1988.

[6] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine". Computer Networks and ISDN Systems. vol. 30, Apr. 1998, pp. 107-117, doi: 10.1016/S0169-7552(98)00110-X.

[7] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", Journal of the ACM, vol. 46, Sept. 1999, pp. 604-632, doi: 10.1145/324133.324140.

[8] S. Feng, L. Zhang, Y. Xiong, and C. Yao, "Focused Crawling Using Navigational Rank", Proc. 19th ACM International Conference on Information and Knowledge Management (CIKM '10), 2010, pp. 1513-1516, doi: 10.1145/1871437.1871660.

[9] G. Almpanidis, C. Kotropoulos, and T. Pitas, "Combining Text and Link Analysis for Focused Crawling-An Application for Vertical Search Engines", Information System, Elsevier, vol. 32, Sept. 2007, pp. 886-908, doi: 10.1016/j.is.2006.09.004.

[10] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur, "The shark-Search Algorithm. An Application: Tailored Web Site Mapping", Computer Networks and ISDN Systems, Elsevier, vol. 30, Apr. 1998, pp. 317-326, doi: S0169-7552(98)00038-5.

[11] X. Chen, X. Zhang, "HAWK: A Focused Crawler with Content and Link Analysis", Proc. IEEE International Conference on e-Business Engineering (ICEBE'08), IEEE Computer Society, Oct. 2008, pp. 677-680, doi: 10.1109/ICEBE.2008.46.

[12] C. Yin, J. Liu, C. Yang, and H. Zhang, "A Novel Method for Crawler in Domain-specific Search", Computational Information System, vol. 5, Dec. 2009, pp. 1749-1755.

[13] A. Pesaranghader and N. Mustapha, "Term Frequency-Information Content for Focused Crawling to Predict Relevant Web Pages", International Journal of Digital Content Technology and its Applications (JDCTA), vol. 7, no. 12, Aug. 2013, pp. 113-122.

[14] A. Pesaranghader, A. Pesaranghader, N. Mustapha, and N. M. Sharef, "Improving Multi-term Topics Focused Crawling by Introducing Term Frequency-Information Content (TF-IC) Measure", 3rd International Conference on Research and Innovation in Information Systems, 2013, pp. 102-106, doi: 10.1109/ICRIIS.2013.6716693.

[15] M. Yuvarani, N.C.S.N. Iyengar, and A. Kannan, "LSCrawler: a framework for an enhanced focused web crawler based on link semantics", Proc. IEEE/WIC/ACM International Conference on Web Intelligence, Dec. 2006, pp. 794-800, doi: 10.1109/WI.2006.112.

[16] S. Ganesh, M. Jayaraj, V. Kalyan, and G. Aghila, “Ontology-based Web Crawler”. Proc. The International Conference on Information Technology: Coding and Computing (ITCC 2004), Apr. 2004, pp. 337-341, doi: 10.1109/ITCC.2004.1286658.

[17] A. B. Can and N. Baykal, “MedicoPort: a Medical Search Engine for All”, Computer Methods and Programs in Biomedicine, Apr. 2007; vol. 86, no. 1, pp. 73-86, doi: 10.1016/j.cmpb.2007.01.007.

[18] A. Pesaranghader, A. Pesaranghader, A. Rezaei, and D. Davoodi, “Gene Functional Similarity Analysis by Definition-based Semantic Similarity Measurement of GO Terms”, Proc. 27th Canadian Conference on Artificial Intelligence, May 2014, pp. 203-214, doi: 10.1007/978-3-319-06483-3_18.

[19] A. Pesaranghader, A. Pesaranghader, and N. Mustapha, “Word Sense Disambiguation for Biomedical Text Mining Using Definition-Based Semantic Relatedness and Similarity Measures”, International Journal of Bioscience, Biochemistry and Bioinformatics, vol. 4, no. 4, 2014, pp. 280-283, doi: 10.7763/IJBBB.2014.V4.356.

[20] A. Pesaranghader, A. Rezaei, and A. Pesaranghader, “Adapting Gloss Vector Semantic Relatedness Measure for Semantic Similarity Estimation: An Evaluation in the Biomedical Domain”, Proc. 3rd Joint International Semantic Technology (JIST'13), Nov. 2013, pp. 129-145, doi: 10.1007/978-3-319-06826-8_11.

[21] J. Caviedes and J. Cimino, “Towards the Development of a Conceptual Distance Metric for the UMLS”, Journal of Biomedical Informatics, Apr. 2004, vol. 37, no. 2, pp. 77-85.

[22] R. Rada, H. Mili, E. Bicknell and M. Blettner, “Development and Application of a Metric on Semantic Nets”, IEEE Transactions on Systems, Man and Cybernetics, Jan-Feb 1989, vol. 19, no. 1, pp. 17-30, doi: 10.1109/21.24528.

[23] Z. Wu and M. Palmer, “Verb Semantics and Lexical Selections”, Proc. 32nd Annual Meeting of the Association for Computational Linguistics, 1994, pp. 133-138, doi: 10.3115/981732.981751.

[24] C. Leacock and M. Chodorow, “Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet”, MIT Press, 1998, pp. 265-283.

[25] J. Zhong, H. Zhu, J. Li and Y. Yu, “Conceptual Graph Matching for Semantic Search”, Proc. 10th International Conference on Conceptual Structures, 2002, pp. 92.

[26] P. Resnik, “Using Information Content to Evaluate Semantic Similarity in a Taxonomy”, Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI), pp. 448-453, vol. 5, Canada, 1995.

[27] J.J. Jiang and D.W. Conrath, “Semantic Similarity based on Corpus Statistics and Lexical Taxonomy”, Proc. International Conference on Research in Computational Linguistics, 1997, pp. 19-33.

[28] D. Lin, “An Information-theoretic Definition of Similarity”, Proc. 15th International Conference on Machine Learning, 1998, pp. 296-304.