Measuring Diversity of a Domain-Specific Crawl

Pattisapu Nikhil Priyatam, Ajay Dubey, Krish Perumal, Dharmesh Kakadia, and Vasudeva Varma

Search and Information Extraction Lab, IIIT-Hyderabad, India
{nikhil.pattisapu, ajay.dubey, krish.perumal, dharmesh.kakadia}@research.iiit.ac.in, vv@iiit.ac.in

Abstract. In this work we present various metrics to measure diversity of a domain-specific crawl. We evaluate these metrics using domain-specific crawl originated from ODP URLs and find that these metrics are indeed able to capture diversity. We argue that these metrics can be used for comparing seed sets and crawling strategies with respect to diversity.

1 Introduction

With exponentially increasing content on Internet the use of domain specific search engines are on rise [13]. Several researchers have argued that using focused crawlers for building domain specific search engines is more efficient [5] [11]. Traditionally, a focused crawler is judged based on its ability to fetch relevant (to a domain) documents. While document relevancy is an important aspect of retrieval, crawl diversity of content is equally important factor that impacts the quality of a domain specific search engine. To a large extent, crawl diversity depends on the focused crawler of a search engine. The efficiency of a focused crawler in turn depends on the choice of seed URLs [12] (the list of URLs that the crawler starts with) and the crawling strategy it uses. Though significant effort has gone into building various focused crawlers, not enough research has been done in evaluating them. Even the very few ones which evaluate focused crawlers rely on metrics related to precision, harvest ratio (rate of change of precision), crawl robustness, etc. To the best of our knowledge, there is no work which discusses metrics to measure diversity of a domain-specific crawl.

In this work we propose various metrics to measure diversity of a **domain-specific** web crawl. To achieve this, we use four different methods: semantic distance, statistical dispersion, average similarity and topic modeling. These metrics can be used to better analyze and compare different seed sets and focused crawling strategies as shown in figure 1. In figure 1, we say Seed set S1 is more diverse than Seed set S2, if the diversity score of crawl C1 is strictly greater than that of C2. The same applies for comparing focused crawlers as well.

2 Related Work

The notion of diversity has received great attention in the problem of search results diversification [8][4][18][16][14]. The problem of maximizing search results

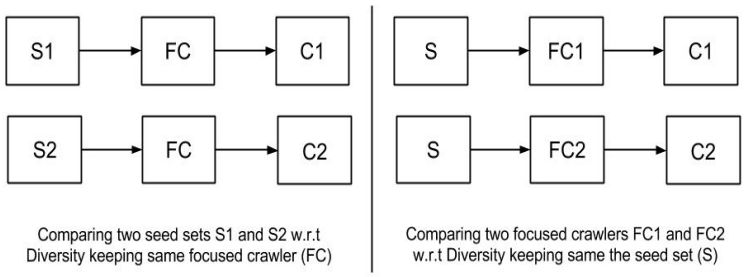

Fig. 1. Comparing Seed sets and Focused Crawlers based on crawl diversity

coverage with respect to different aspects of a query is NP-hard [1]. Most previous works on search result diversification are based on a greedy approximation to this problem [17].

Santos et al. use a sub-query based method for search results diversification [17][15]. They assume queries submitted to a retrieval system are ambiguous. Based on this assumption they submit several sub-queries to the retrieval system wherein each sub-query captures a different "aspect" of the query. They present to the user the merged list of ranked retrieved documents.

Dou et al. [3] argue that search results should be diversified in a multi-dimensional way, as queries are usually ambiguous at different levels and dimensions. They mine subtopics from anchor texts, query logs, search result clusters, and web sites and propose techniques to diversify search results based on multiple dimensions of subtopics. They claim that, by incorporating multiple types of subtopics, their models improve the diversity of search results over the sole use of one of them.

Apart from search results, diversity also plays a crucial role in recommender systems. Zhou et al. [19] mention that the key challenge in making useful recommendation is that while the most useful individual recommendations are to be found among diverse niche objects, the most reliably accurate results are obtained by using user or object similarity. They come up with a hybrid approach to resolve this dilemma. They use two features to judge the diversity, which they call 'personalization' (inter-user diversity) and 'surprisal/novelty' (capacity of the recommender system to generate novel and unexpected results to suggest objects a user is unlikely to know about already).

As discussed above, there are several metrics to evaluate retrieval systems and recommender systems w.r.t diversity, but there is no single work which evaluates a focused crawler w.r.t diversity. Even the very few ones which evaluate focused crawlers rely on metrics related to precision, harvest ratio (rate of change of precision), crawl robustness, etc. Menczer et al. [7] propose various methods for evaluating topic specific crawl. In the *Assessment via Classifiers* method, they train a classifier for each topic and evaluate the precision of the crawled set. This requires huge amount of accurate training data (manual tagging) which is

labour intensive. The second method *assessment via a retrieval system* is based on the intuition that a crawler should retrieve good pages earlier than the bad ones. The last method *Mean Topic Similarity* measures the cohesiveness of the crawled set with the topic as the core. The underlying assumption is that the more cohesive the crawled set the more relevant its pages. To the best of our knowledge, there is no work which discusses metrics to evaluate a domain specific crawl.

3 Approach

In this section we present various approaches to measure the diversity of a given web crawl. Each web page is represented in the form of a text document which contains the parsed text of the web page. In the rest of this paper we refer to a web page as a document.

3.1 Semantic Distance

This measure uses semantic distance of documents within a crawl to calculate its diversity. The main intuition behind this metric is that a more diverse web crawl will have a higher semantic distance between its documents when compared to a less diverse web crawl. Semantic distance between two documents is defined as the average Wordnet distance between their top k keywords (described in equation 1). The crawl diversity(D1 score) is then computed as the average semantic distance between every document pair as shown in equation 2, where N represents the total number of documents present in the crawl. The Wordnet distance is calculated using Wordnet similarity as used by Pedersen et al. [10]. The Wordnet distance function is designed in a way that highly similar words (or same words) get a score 0 and highly dissimilar words get a score of 1.

$$SD(d_x, d_y) = \sum_{i=1}^{k} \sum_{j=1}^{k} Wordnet\ Distance(w_{xi}, w_{y_j})^{1} \tag{1}$$

where $SD(d_x, d_y)$ represent the semantic distance between documents x and y respectively and w_{xi}, w_{y_j} represents i^{th} word of document x and j^{th} word of document y respectively.

$$D1\ Score\ = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} SD(d_i, d_j)}{N^2} \tag{2}$$

[1] `http://rednoise.org/rita/wordnet/documentation/riwordnet_method_getdistance.htm`

3.2 Dispersion around a single centroid

Dispersion refers to the spread or variability in a variable. We measure the variance across the crawled set of documents to judge its diversity. The D2 score or variance, as shown in equation 3, is calculated as the average squared distance of all documents from the mean. Here $\boldsymbol{d_i}$ refers to i^{th} document represented as a vector, $\boldsymbol{\mu}$ represents the mean of all $\boldsymbol{d_i}$ $'s$ and N represents total number of documents.

$$D2\ Score\ = \frac{\sum_{i=1}^{N}(\boldsymbol{d_i} - \boldsymbol{\mu})^2}{N} \tag{3}$$

In the above equation we have represented each document as a vector. We consider two feature spaces i.e. bag of words and context vectors.

Bag of Words In this model each document is represented as bag-of-words or a vector of words over entire vocabulary. The value corresponding to each word is its frequency within the document.

Context Vectors Since the bag of words model suffers from the problem of sparsity, we also use the context vector model. The context vector is a much more compact representation of a document, where the document is represented as a centroid of the top n word vectors. Figure 2 shows the context vectors of two documents d_A, d_B (shown in red) with word vectors w_{A1}, w_{A2} and w_{B1}, w_{B2} (shown in black) respectively. The word vectors of both the documents are projected onto a common feature space consisting of words occurring in both the documents. Figure 2 depicts the common space as containing the words i and j. A word vector is a vector of k words surrounding it, where k can be understood as the window size. Again, the value corresponding to each word is its frequency within the document. The concept of context vectors is explained in detail in [9].

3.3 Average Similarity between Document Pairs

In this metric, diversity is measured using the average cosine similarity between every pair of documents. This is shown in the crawl shown in equation 4. The crawl diversity is then calculated as shown in equation 5. The intuition behind this metric is that, higher the similarity between the documents, lesser is the diversity of the crawl. We use two document representations of a document in this metric as well. The diversity is calculated as the inverse of average cosine similarity.

$$ACS = \frac{\sum_{i=1}^{k}\sum_{j=1,\ j\neq i}^{k} Cosine\ Similarity(d_i, d_j)}{number\ of\ document\ pairs} \tag{4}$$

$$D3\ Score = \frac{1}{ACS} \tag{5}$$

3.4 Using Topic Models

In this metric, we run LDA [2] on n documents randomly picked from the crawl to get k topics. The D4 score is calculated as the sum of KL divergence [6] values between every two topics as shown in equation 6.

$$D4\ Score = \sum_{i=1}^{k} \sum_{j=1,\ j \neq i}^{k} KLDivergence(t_i, t_j) \tag{6}$$

where t_i and t_j represent topic i and topic j, and

$$KLDivergence(t_i, t_j)\ = \sum_{v=1}^{|V|} ln\left(\frac{t_i(v)}{t_j(v)}\right) t_i(v) \tag{7}$$

where $t_i(v)$ and $t_j(v)$ represent the probabilities of word v in topics i and j respectively and $|V|$ represents vocabulary size. Hence a web crawl covering varied topics will have a higher diversity score than the crawl containing similar topics.

4 Evaluation Methodology

In this section we validate the metrics discussed in section 3. We argue that a crawl diversity metric is valid if and only if, it satisfies the constraint: *diversity score of a web crawl, generated by a diverse seed URL set is strictly greater than diversity score of a web crawl, generated by a less diverse seed URL set.* For generating a diverse and a less diverse domain-specific crawl, we selectively pick two sets of URLs of same size from ODP. The first set contains URLs of a wide range of topics of a particular domain and the second set contains URLs from the subset of these topics. For instance, in case of a *Health* specific crawl a set of ODP URLs under the top level categories - *aging*, *fitness*, *nutrition*, *insurance*, etc. would correspond to a diverse set of URLs. ODP URLs under the deeper level category of *dietitian* would correspond to a less diverse set of URLs. This is illustrated in figure 3. We crawl these two sets of URLs thus giving rise to more diverse and less diverse web crawls.

5 Experimental Setup

For this work, we experiment on three domains - tourism, health and sports. For tourism, we pick lodging as the corresponding less diverse seed set. Similarly, for health and sports, we pick dietitian and badminton respectively. 200 URLs are picked from each of these to be used as seed URLs, and a depth 1 crawl is

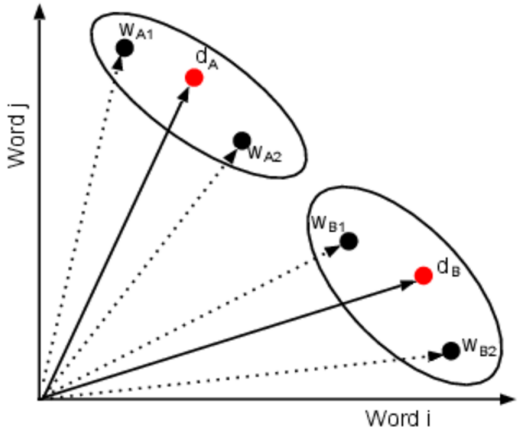

Fig. 2. Context Vectors

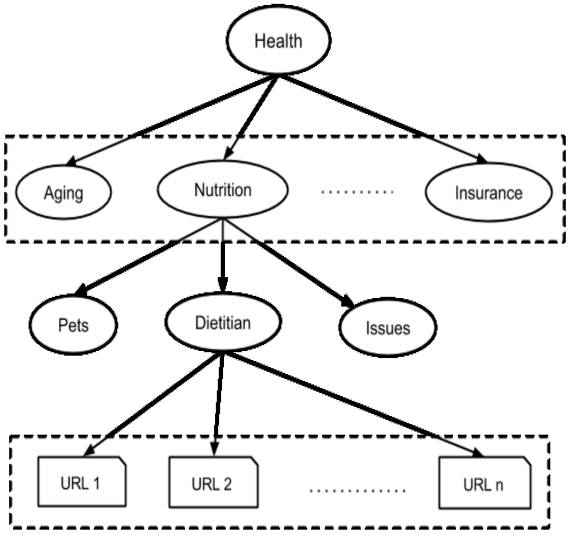

Fig. 3. Picking URLs from ODP Hierarchy

performed. These URLs and their respective crawls have been publicly released for research purposes [2]. In the methods using context vectors, we use 5 context vectors per document and word vectors are generated using a window of 4 surrounding words. In case of the topic modeling approach, we use number of topics as 5. The hyper-parameters for LDA are α is 50 and β is 0.01.

[2] `https://www.dropbox.com/sh/i66hzq5cu9aq50k/0_idXZ5s0L`

6 Results

Tables 1, 2, 3 and 4 show the diversity scores for the metrics based on semantic distance, topic modeling, average similarity between document pairs and dispersion respectively. MD refers to the more diverse crawl (i.e. crawl originated from diverse seed set) and LD refers to the less diverse crawl (Crawl originated from less diverse seed set). The term "Ratio" refers to the ratio of diversity score of MD to diversity score of LD.

	MD	LD	Ratio
Tourism	0.788	0.779	1.01
Health	0.759	0.721	1.05
Sports	0.764	0.769	0.99

Table 1. Semantic Distance Based Metric

	MD	LD	Ratio
Tourism	4.475	4.470	1.001
Health	4.645	4.406	1.054
Sports	4.470	4.420	1.011

Table 2. Topic Modeling Based Approach

Feature Space	Bag of Words			Context Vectors		
	MD	LD	Ratio	MD	LD	Ratio
Tourism	53.19	31.35	1.69	83.33	46.26	1.77
Health	49.54	18.61	2.66	83.34	26.52	2.91
Sports	37.60	14.07	2.67	55.55	18.21	3.04

Table 3. Similarity Based Metric

Feature Space	Bag of Words			Context Vectors		
	MD	LD	Ratio	MD	LD	Ratio
Tourism	118.31	109.42	1.08	38.37	36.82	1.04
Health	169.58	110.56	1.53	52.79	37.07	1.42
Sports	118.55	100.48	1.17	40.70	36.21	1.12

Table 4. Dispersion Based Metric

7 Analysis and Insights

We observe that the diversity measure based on average similarity between document pairs outperforms the rest of the approaches. We also find that the semantic distance metric fails to distinguish diverse and less diverse crawls. In fact, it wrongly identified a diverse sports crawl as being less diverse and vice versa i.e. $\frac{D(MD_{sports})}{D(LD_{sports})} < 1$. Upon deeper inspection, we find that many words and

their spell variations are not present in Wordnet. This has adversely affected the performance of the metric. Also, we find that our crawl contains ill-parsed and non-English language documents because of which meaningful topics were not formed, thereby affecting the performance of the topic modeling based approach.

8 Conclusions and Future Work

This work presents four metrics to measure diversity of a domain-specific crawl which are useful in the context of domain-specific search engines. We rank these metrics based on their ability to differentiate between crawls originated from diverse and less diverse seeds. From the experiments on the tourism, health and sports domains, we observe that the cosine similarity based metric outperforms all others. In both cosine similarity and dispersion based measures context vectors proved to be a better feature space than bag of words. All the proposed metrics, except the semantic distance based metric, are language independent. Even, the semantic distance metric can be easily extended to other languages for which a concept hierarchy like Wordnet is available.

In future, we would like to work on the relevant and correctly parsed portions of the crawl with the help of state of the art parsers and classifiers. The proposed metrics can be used to better analyze and compare different focused crawling strategies. Moreover, more efficient focused crawlers can be built by analyzing the diversity of the previously crawled content, thus leading to higher diversity of the resultant domain-specific crawl. The current work does not use domain knowledge or any external resource to evaluate the crawl diversity. In future, we wish to use the subtopic structure of a domain to evaluate crawl diversity.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.
2. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
3. Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 475–484. ACM, 2011.
4. M. Drosou and E. Pitoura. Search result diversification. *ACM SIGMOD Record*, 39(1):41–47, 2010.
5. D. Fesenmaier, H. Werthner, and K. Wober. Domain specific search engines. In *Travel Destination Recommendation Systems: Behavioural Foundations and Applications*, pages 205–211. CABI, 2006.
6. S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
7. F. Menczer, G. Pant, P. Srinivasan, and M. Ruiz. Evaluating topic-driven web crawlers. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 241–249. ACM, 2001.

8. E. Minack, G. Demartini, and W. Nejdl. Current approaches to search result diversification. In *Proc. of 1st Intl. Workshop on Living Web*. Citeseer, 2009.
9. S. Patwardhan and T. Pedersen. Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*, volume 1501, pages 1–8, 2006.
10. T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics, 2004.
11. N. Priyatam, V. Reddy, and V. Varma. Domain specific search in indian languages. In *Proceedings of the First Workshop on Information and Knowledge Management for Developing Regions (IKM4DR) at CIKM 2012*, pages 23–30. ACM, 2012.
12. P. N. Priyatam, A. Dubey, K. Perumal, S. Praneeth, D. Kakadia, and V. Varma. Seed selection for domain-specific search. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 923–928. International World Wide Web Conferences Steering Committee, 2014.
13. P. N. Priyatam, S. Iyengar, K. Perumal, and V. Varma. Don't use a lot when little will do: Genre identification using urls. *Research in Computing Science*, 70:207–218, 2013.
14. F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 691–692. ACM, 2006.
15. R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890. ACM, 2010.
16. R. L. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. In *Advances in Information Retrieval Theory*, pages 250–261. Springer, 2011.
17. R. L. Santos, J. Peng, C. Macdonald, and I. Ounis. Explicit search result diversification through sub-queries. In *Advances in information retrieval*, pages 87–99. Springer, 2010.
18. R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proceedings of the 18th international conference on World wide web*, pages 341–350. ACM, 2009.
19. T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.