International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 3 Issue 2, February-2014, pp: (39-43), Impact Factor: 1.147, Available online at: www.erpublications.com

Language Model Issues in Web Spam Detection

K. Priya Karunakaran[1], Seema Kolkur[2]
[1]Computer Engineering Department, St. Francis Institute of Technology, Mumbai, India
[2]Computer Engineering Department, Thadomal Shahani Engineering College, Mumbai, India

Abstract: Language models have been widely used in the detection of spam pages in the web. However even though most of the experiments using language models to detect spam have got improved results, there exists several problems in the use of language models which affects the validity of the results. This paper points out the shortcomings of using language models specifically KL-Divergence and suggested improvements.

Keywords: Jansen- Shannon Divergence, KL- Divergence, Language model, Web spam detection.

Introduction

Web spam is a serious problem for search engines because it strongly degrades the quality of the results. Web spam involves all the techniques used for the purpose of getting an undeservedly high rank. Generally, there are three types of Web spam: link spam, content spam, and cloaking, a technique in which the content presented to the search engine spider is different to that presented to the browser of the user. One of the most successful techniques for Web spam detection, as it can be seen in the AIRWeb competition1, is the definition of features which take different values for spam and non-spam pages. These features are thus used to implement a classifier able to detect spam pages.

To improve web spam detection, Juan, Lourdes et al. [2] proposed a technique that checks the coherence between a page and one pointed by any of its links. Two pages linked by a hyperlink should be semantically related, by at least a weak contextual relation. They make a Language Model from each source of information and ask how different these two language models are from each other. These sources of information are: i) anchor text, surrounding anchor text and URL terms from the source page, and ii) title and content from the target page [3]. They apply Kullback-Leibler (KL) divergence [7] on the language models to characterize the relationship between two linked pages. The result is a system that significantly improves the detection of Web spam using fewer features. However this paper makes the observation that compared to classifiers generated using features other than language model based features the results obtained by a classifier using just language model features is less than that achieved by using other features such as content based and link based features. Using language models as a way of identifying spam and non-spam web pages is a very important technique compared to other approaches as a language model gives a logical view of a web page.

A web page is considered to be spam or not spam by a viewer basically by his expectations about that page and what he gets to view in that page. Thus language modeling technique to detect web spam is a very promising technique and needs to be improved. This paper has tried to identify the causes of the poor performance of language model approach as shown in the results obtained by Juan, Lourdes et al. [2] and tries to give alternative solutions.

Language Models

Language models are probabilistic methods that have been previously used successfully in areas of speech recognition, machine translation, part-of-speech tagging, parsing and information retrieval. Statistical language models have been developed to capture linguistic features hidden in texts, such as the probability of words or word sequences in a language. A statistical language model (SLM) is a probability distribution P(s) over strings S that attempts to reflect how frequently a string S occurs as a sentence. Previous works have proved that language model disagreement techniques are very efficient in tasks such as blocking blog spam [10] or detecting nepotistic links [11].

J. M. Ponte and W. B. Croft et.al [6] suggests that for coming up with good queries is to think of words that would likely appear in a relevant document, and to use those words as the query. The language modelling approach to IR directly models that idea: a document is a good match to a query if the document model is likely to generate the query, which will in turn happen if the document contains the query words often. The most popular divergence measures used are KL- divergence, Pearson divergence [1], the relative Pearson divergence, and the L2-distance.

International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 3 Issue 2, February-2014, pp: (39-43), Impact Factor: 1.147, Available online at: www.erpublications.com

Kullback-Leibler Divergence

Kullback-Leibler Divergence (KLD) [9] is one of the most successful methods based on term distribution analysis to compute the divergence between the probability distributions of terms of two documents. KLD of two text units T1 and T2 is computed as follows:

$$\mathrm{KLD}(T_1||T_2) = \sum_{t \in T_1} P_{T_1}(t) \log \frac{P_{T_1}(t)}{P_{T_2}(t)} \quad (1)$$

where PT1 (t) is the probability of the term t in the first text unit, and PT2 (t) is the probability of the term t in the second text unit.

KL divergence is an asymmetric divergence measure. That is, KLD(T1||T2) ≠ KLD(T2||T1). One disadvantage of using KL divergence as a ranking function is that scores are not comparable across queries. This problem does not affect ad-hoc retrieval, but is important in other applications such as topic tracking [5].

A. Limitations of KL-Divergence

Consider KL measures the divergence between a probability distribution P and Q.

Think of P as (x_1: p_1, x_2: p_2, ..., x_n:p_n) with sum (p_i) = 1.0 and Q as (y_1:q_1, ..., y_n:q_n) with sum(q_j) = 1.0. Assume for now that P and Q are defined over the same outcomes x_i. Then the definition of KL is:

$$KL(P,Q) = \mathrm{sum}(i = 1..n)\ [p\_i * \log(p\_i/q\_i)] \quad (2)$$

When we try to compute this formula, we must address two questions:

What to do if qi = 0 for some i or pi = 0 for some i?

How do we define the formula when P and Q are defined over different samples?

The direct answer is that we always consider that: 0 * log(0) = 0.

This handles the case of pi=0, and when both pi and qi are 0 (since log(pi/qi) = log(pi) - log(qi)). But not the case when pi!=0 and qi=0. The general definition is that in such a case, the divergence is infinite. This means that if one distribution predicts that an event is possible (p(e)>0) and the other predicts it is absolutely impossible (q(e)=0), then the two distributions are absolutely different.

However, P and Q are derived from observations and sample counting -- that is, P and Q are probability distributions derived from frequency distributions. In this case, the solution is that we should never predict in the derived probability distribution that an event is completely impossible: when we derive the probability distribution, we must take into account the possibility of unseen events. Since there are infinitely many unseen events, this makes it difficult to compare frequency distributions that predict non-zero probability for unseen events.

B. KL-Divergence of two documents

Alberto, Paolo, and Jose-Miguel et. al [4] has attempted to demonstrate the problem of finding the KLD for two small documents where we may end up with an empty set of terms thereby not being able to generate the KLD.

Let P and Q be two probability distributions of a discrete random variable. If the following two properties hold:

when P and Q both sum to 1and for any i such that P(i) > 0 and Q(i) > 0; then, we can define their KL-divergence as:

$$D_{KL}(P||Q) = \Sigma i\ P(i) \log(P(i)/Q(i)) \quad (3)$$

and it has three properties:

1) DKL (P||Q) ≠ DKL(Q||P) (asymmetry)

2) It is additive for independent distributions

3) DKL ≥ 0 with DKL= 0 iff P = Q

We regard a document d as discrete distribution of |d| random variables, where |d| is the number of words in the document. Now, let d1 and d2 be two documents for which we want to calculate their KL-divergence. We run into two problems:

We need to compute the KL-divergence twice due to asymmetry: DKL (d1||d2) and DKL(d2||d1).

Due to the 2nd constraint for defining KL-divergence, our calculations should only consider words occurring in both d1 and d2.

To illustrate the problem of handling documents with no or little overlapping vocabularies, consider the following documents:

d1: This is a document

d2: This is a sentence

After removing the stop words (this, is, a) we get:

d1: document

d2: sentence

According to constraint 2, we need to operate on the intersection of the documents' vocabularies:

d1 ∩ d2 = ∅

We end up with the empty set and therefore we cannot compute directly the KL-divergence. In this case we can assign it a large number like 1e33.

When we have larger documents, for example:

d1: Many research publications want you to use BibTex, which better organizes the whole process. Suppose for concreteness your source file x.tex. Basically, you create a file x.bib containing the bibliography, and run bibtex on that file.

d2: In this case you must supply both a \left and a \right because the delimiter height are made to match whatever is contained between the two commands. But, the \left doesn't have to be an actual 'left delimiter', that is you can use '\left)' if there were some reason to do it.

After stop-word removal, lowercasing and discarding words less than 2 characters, the documents become:

d1: many research publications want you use bibtex better organizes whole process suppose concreteness your source file tex basically you create file bib containing bibliography run bibtex file

d2: case you must supply both left right because delimiter height made match whatever contained between two commands left doesn't have actual left delimiter you use left some reason

The vocabulary intersection of the documents consists of two terms: "use" and "you". In d1 "use" occurs 1 time and "you" occurs 2 times. Surprisingly, in d2 "use" also occurs 1 time and "you" occurs 2 times too. The distributions d1and d2are equal, and therefore DKL(d1||d2)=0. So these documents are deemed equal. A better stop-word list could have removed "use" and "you" and in that case the documents would have an infinite KL-divergence as in the first example. However it is easy to think of similar examples where stop-word lists wouldn't have been of much help [4].

This shows how finding the KL-divergence between a web page and its linked page may not necessarily give the expected output.

Jensen-Shannon Divergence

Jensen-Shannon divergence (JSD) can be defined as measure of the "distance" or similarity between two probability distributions [8]. It can also be generalized to measure the distance between a finite number of distributions.

JSD is a natural extension of the KLD that can be applied to a set of distributions [9]. KLD can be defined between two distributions, while the JSD of a set of distributions is the average KLD of each distribution to the mean of the set. Unlike KLD, JSD is a true metric and is bounded. If a classifier can provide a distribution of class membership probabilities for a given example, then we can use JSD to compute a measure of similarity between the distributions produced by a set (ensemble) of such classifiers.

If Pi(x) is the class probability distribution given by the i-th classifier for the example x (abbreviated as Pi) we can then compute the JSD of a set of size n as:

$$JS(P1, P2, ..., Pn) = H(\Sigma n\ i=1\ \ wiPi) - \Sigma n\ i=1\ \ wi\ H(Pi) \quad (4)$$

where wi is the vote weight of the i-th classifier in the set and H(P) is the Shannon entropy of the distribution:

P = {pj : j = 1, ...,K}, defined as,

$$H(P) = -\Sigma K\ j=1\ \ pj \log pj \quad (5)$$

Higher values for JSD indicate a greater spread in the class probability estimates distributions, and it is zero if and only if the distributions are identical. JSD can be used to measure the utility of examples in active learning for improving classification accuracy.

As an extension of KL-Divergence we can compute JS-Divergence of two probability distributions P and Q as:

$$JSD(P||Q)=1/2\ KLD(P1||M)+1/2KLD(Q||M) \quad (6)$$

where, M=1/2 (P+Q).

Consider again the two documents d1 and d2:

d1: many research publications want you use bibtex better organizes whole process suppose concreteness your source file tex basically you create file bib containing bibliography run bibtex file

d2: case you must supply both left right because delimiter height made match whatever contained between two commands left doesn't have actual left delimiter you use left some reason

By calculating the JSD between d1 and d2 we get: JSD (d1||d2) = 0.2093 which shows that the two documents are more similar and does not give an infinite value as given by KLD.

Conclusion

KL-Divergence suffers from two drawbacks: It is not symmetric in its arguments and it does not naturally generalize to measuring the divergence among documents with little or no overlapping vocabularies. Jensen-Shannon Divergence can overcome the limitations of KL-Divergence in finding the similarities between a web page and its linked pages due to its symmetry and generalization properties. Thus we suggest Jenson – Shannon be used instead of KL-Divergence in the detection of web spam to improve results in projects that involve the generation of language models for detecting spam in web pages.

References

[1]. Masashi Sugiyama, Song Liu, and Marthinus Christoffel du Plessis, "Direct Divergence Approximation between Probability Distributions and Its Applications in Machine Learning", Journal of Computing Science and Engineering, Vol. 7, No. 2, June 2013, pp. 99-111

[2]. Lourdes Araujo, Juan Martinez-Romo, "Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models". IEEE Transactions On Information Forensics And Security, Vol. 5, No. 3, September 2010

[3]. K. Priya Karunakaran and Seema Kolkur, "Review of Web Spam Detection Techniques", International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 2 Issue 4 July 2013

[4]. Alberto Barron-Cedeno, Paolo Rosso, and Jose-Miguel Benedi, , "Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance", CICLing 2009, LNCS 5449, pp. 523–534, 2009

[5]. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[6]. J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in Proc. 21st Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'98), New York, 1998, pp. 275–281, ACM.

[7]. S. Kullback and R. A. Leibler, "On Information and sufficiency", The annals of Mathematical Statistics, vol. 22, no. 1, pp. 79-86, 1951.

[8]. Prem Melville, Stewart M. Yang, Maytal Saar-Tsechansky and Raymond Mooney "Active Learning for Probability Estimation using Jensen-Shannon Divergence", Proceedings of The 16th European Conference on Machine Learning (ECML),Porto, Portugal, pp. 268-279, October 2005

[9]. T. M. Cover and J. A. Thomas. Elements of information theory. Wiley-Interscience, New York, NY, USA, 1991.

[10]. G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," in Proc. First Int. Workshop on Adversarial Information Retrieval on theWeb (AIRWeb), Chiba, Japan, 2005, pp. 1–6.

[11]. A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uher, "Detecting nepotistic links by language model disagreement," in Proc. 15th Int. Conf. World Wide Web (WWW'06), New York, 2006, pp. 939–940, ACM.