WIKIPEDIA

Divergence-from-randomness model

In the field of information retrieval, **divergence from randomness**,one of the very first models, is one type of probabilistic model. It is basically used to test the amount of information carried in the documents. It is based on Harter's 2-Poisson indexing-model. The 2-Poisson model has a hypothesis that the level of the documents is related to a set of documents which contains words occur relatively greater than the rest of the documents.It is not really a 'model', but a framework for weighting terms using probabilistic methods, and it has a special relationship for Term weighting based on notion of eliteness.

Term weights are being treated as the standard of whether a specific word is in that set or not. Term weights are computed by measuring the divergence between a term distribution produced by a random process and the actual term distribution.

DFR models set up by instantiating the three main components of the framework: first selecting a basic randomness model, then applying the first normalization and at last normalizing the term frequencies. The basic models are from the following tables.

Contents

Definition

The divergence from randomness is based on this idea: "The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in document d. In other words, the term-weight is inversely related to the probability of term-frequency within the document d obtained by a model M of randomness."[1](By terrier.org)

$$\mathbf{weight}(t|d) = k\mathbf{Prob}_M(t \in d|\mathbf{Collection})$$ (Formula 1)

1. M represents the type of model of randomness which employs to calculate the probability.
2. d is the total amount of words in the documents.

3. t is the amount of a specific word in d.
4. k is defined by M.

It is possible that we use different urn models to choose the appropriate model M of randomness. In Information Retrieval, we have documents instead of urns, and terms instead of colors. There are several ways to choose M, each of these has a basic DFR model to support it.

Model

Basic DFR Models

```
D      Divergence approximation of the binomial
P      Approximation of the binomial
BE        Bose-Einstein distribution
G  Geometric approximation of the  Bose-Einstein
I(n)   Inverse Document Frequency Model
I(F)   Inverse Term Frequency Model
I(ne) Inverse Expected Document Frequency Model
```

DFR Models

```
BB2    Bernoulli-Einstein model with Bernoulli after-effect and normalization 2.
IFB2    Inverse Term Frequency model with Bernoulli after-effect and normalization 2.
In-expB2 Inverse Expected Document Frequency model with Bernoulli after-effect and normalization 2. The logarithms are
base 2,.This model    can be used for classic ad-hoc tasks.
In-expC2  Inverse Expected Document Frequency model with Bernoulli after-effect and normalization 2. The logarithms are
base e. This model can be used for classic ad=hoc tasks.
InL2    Inverse Document Frequency model with Laplace after-effect and normalization 2. This model can be used for tasks
that require early precision.
PL2    Poisson model with Laplace after-effect and normalization 2. This model can be used for tasks that require early
precision[7,8].
```

First Normalization

When a specific a rare term cannot be found in a document, then in that document the term has approximately zero probability of being informative. On the other hand, if a rare term has a lot of occurrences in a document, therefore it can have a very high, closed to 100%, probability to be informative for the topic that mentioned by the document. Applying to Ponte and Croft's language model can also be a good idea.Noticed that a risk component is considered in the DFR. Logically speaking, if the term-frequency in the document is relatively high, then inversely the risk for the term of not being informative is relatively small. Say we have a Formula 1 giving a high value, then a minimal risk has the negative effect of showing small information gain. So we choose to organize the weight of Formula 1 to only consider the portion of which is the amount of information gained with the term. The more the term occurs in the elite set, the less term-frequency is due to randomness, and thus the smaller the associated risk is. We basically apply two models to compute the information gain with a term within a document:

```
the Laplace L model, the ratio of two Bernoulli's processes B.
```

Term Frequency Normalization

Before using the within-document frequency tf of a term, the document-length dl is normalized to a standard length sl. Therefore, the term-frequencies tf are recalculated with the respect to the standard document-length, that is:

$$tf_n = tf * \log(1+ sl/dl) \text{ (normalization 1)}$$

tfn represents the normalized term frequency. Another version of the normalization formula is the following:

$$tf_n = tf * \log(1 + c*(sl/dl)) \text{ (normalization 2)}$$

Normalization 2 is usually considered to be more flexible, since you can don't get a fix value for c.

1. tf is the term-frequency of the term t in the document d
2. dl is the document-length.
3. sl is the standard length.

Mathematic and statistical tools

The probability space

Sampling space V

Utility-Theoretic Indexing developed by Cooper and Maron is a theory of indexing based on utility theory. To reflect the value for documents that is expected by the users, index terms are assigned to documents. Also, Utility-Theoretic Indexing is related an “event space” in the statistical word. There are several basic spaces Ω in the Information Retrieval. A really simple basic spaceΩ can be the set V of terms t, which is called the vocabulary of the document collection. Due to Ω=V is the set of all mutually exclusive events, Ω can also be the certain event with probability:

$$P(V)= \sum(t\in V)P(t)=1$$

Thus P, the probability distribution, assigns probabilities to all sets of terms for the vocabulary. Notice that the basic problem of Information Retrieval is to find an estimate for P(t). Estimates are computed on the basis of sampling and the experimental text collection furnishes the samples needed for the estimation. Now we run into the main concern which is how do we treat two arbitrary but heterogeneous pieces of texts appropriately. Paragons like a chapter in a Science Magazine and an article from a sports newspaper as the other. They can be considered as two different samples since those aiming at different population.

Sampling with a document

The relationship of the document with the experiments is made by the way in which the sample space is chosen. In IR, term experiment, or trial, is used here with a technical meaning rather than a common sense. For example, a document could be an experiment which means the document is a sequence of outcomes t∈V, or just a sample of a population. We will talk about the event of observing a number Xt =tf of occurrences of a given word t in a sequence of experiments. In order to introduce this event space, we should introduce the product of the probability spaces associated with the experiments of the sequence. We could introduce our sample space to associate a point with possible configurations of the outcomes. The one-to-one correspondence for sample space can be defined as:

$$\Omega=Vld$$

Where ld is the number of trials of the experiment or in this example, the length of a document. We can assume that each outcome may or may not depend on the outcomes of the previous experiments. If the experiments are designed so that an outcome is influencing the next outcomes, then the probability distribution on V is different at each trial. But, more commonly, in order to establish the simpler case when the probability space is invariant in IR, the term independence assumption is often made. Therefore, all possible configurations ofΩ=Vld are considered equiprobable. Considering this assumption, we can consider each document a Bernoulli process. The probability spaces of the product are invariant and the probability of a given sequence is the product of the probabilities at each trial. Consequently, if p=P(t) is the prior probability that the outcome is t and the number of experiments is ld we obtain the probability of Xt =tf is equal to:

$$P(Xt=tf|p)=(ld\ pick\ tf)p^{tf}q^{ld-tf}$$

Which is the sum of the probability of all possible configurations having tf outcomes out of ld. P(Xt=tf|p) is a probability distribution because

$$\sum(t\in V)P(Xt=tf|p)=(p+q)^{ld}=1$$

1. ld The length of document d.
2. tf The term frequency of t in document d.
3. Xt The number of occurrence of a specific word in one list.

Multiple samplings

Already considering the hypothesis of having a single sample, we need to consider that we have several samples, for example, a collection D of documents. The situation of having a collection of N documents is abstractly equivalent to the scheme of placing a certain number Tot of V colored types of balls in a collection of N cells. For each term t∈V a possible configuration of ball placement satisfies the equations:

$$tf_1+\ldots+tf_N=Ft$$

And the condition

$$F_1+\ldots+F_V=Tot$$

Where Ft is the number of balls of the same color t to be distributed in the N cells. We have thus changed the basic space. The outcome of our experiment will be the documents d in which the ball will be placed. Also, we will have a lot of possible configurations consistent with the number of colored balls.

1. Ft The total number of tokens of t in the collection.
2. Tot The total number of tokens in the collection D

Distributions

Binomial distribution

Hypergeometric Distribution

Bose-Einstein statistics

Fat-tailed distributions

Conclusion

The divergence from Randomness Model is based on the Bernoulli model and its limiting forms, the hypergeometric distribution, Bose-Einstein statistics and its limiting forms, the compound of the binomial distribution with the beta distribution, and the fat-tailed distribution. Divergence from randomness model shows a unifying framework that has the potential constructing a lot of different effective models of IR.

Applications

Applications and Charcteristics

1. The Divergence from randomness model can be applied in automatic indexing in Information Retrieval. These can be explained as the dissertation eliteness, the notion of an informative content of a term within a document.
2. The effectiveness of the models based on divergence from randomness is very high in comparison with both BM25 and language model. For short queries, the performance of the models of divergence from randomness is definitely better than the BM25 Model, which since 1994 has been used as a standard baseline for the comparison of the models.
3. The Divergence from randomness model can show the best performance with only a few documents comparing to other query expansion skills.
4. The framework of Divergence from randomness model is very general and flexible. With the query expansion provided for each component, we can apply different technologies in order to get the best performance.

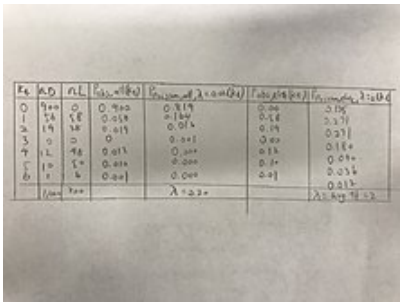
Proximity

Proximity can be handled within DFR to consider the number of occurrences of a pair of query terms within a window of pre-defined size. To specify, the DFR Dependence Score Modifier DSM implements both the pBiL and pBiL2 models, which calculate the randomness divided by the document's length, rather than the statistics of the pair in the corpus the pair in the corpus.

Examples of DFR

1.

Let t be a term and c be a collection. Let the term occur in tfc=nL(t,c)=200 locations, and in df(t,c)=nL(t,c)=100 documents. The expected average term frequency is avgtf(t,c)=200/100=2; this is the average over the documents in which the term occurs. Let N.D(c)=1000 be the total amounts of documents. The term's occurrence is 10% in the documents: P.D(t|c)=100/1000. The expected average term frequency is 200/1000=1/5, and this is the average over all documents. The term frequency is shown as Kt =0,...,6.



The following table show the column nD is the number of Documents that contains kt occurrence of t, shown as nD(t,c,kt). Another column nL is the number of Locations at which the term occurs follows by this equation: nL= kt*nD. The columns to the right show the observed and Poisson probabilities. P obs,elite(Kt) is the observed probability over all documents. P poisson,all,lambda(Kt) is the Poisson probability, where lambda(t,c)=nL(t,c)/N D(c)=0.20 is the Poisson parameter. The table illustrates how the observed probability is different from the Posson probability. P poisson(1) is greater than P obs(1), whereas for kt>1.the observed probabilities are greater than the Poisson probabilities. There is more mass in the tail of the observed distribution than the Poisson distribution assumes. Moreover, the columns to the right illustrate the usage of the elite documents instead of all documents. Here, the single event probability is based on the locations of elite documents only.

Further interest of examples

1. Adjusting documentation length (http://ieomsociety.org/ieom2014/pdfs/513.pdf).
2. Applying DFR in content-only XML Documents (http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Abolhassani_Fuhr_04.pdf)
3. Introduction to DFR models (https://agoldst.github.io/dfrtopics/introduction.html)

References

1. "Divergence From Randomness (DFR) Framework" (http://terrier.org/docs/v2.2.1/dfr_description.html). Terrier Team, University of Glasgow.

- Amati, G. (n.d.). Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness [Abstract]. The university of Glasgow, Fondazione Ugo Bordoni and CORNELIS JOOST VAN RIJSBERGEN University of Glasgow. Retrieved from http://theses.gla.ac.uk/1570/1/2003amatiphd.pdf
- He, B. (2005, April 27). DivergenceFromRandomness. Retrieved from http://ir.dcs.gla.ac.uk/wiki/DivergenceFromRandomness

External links

- Terrier's DFR Web page (http://terrier.org/docs/v3.5/dfr_description.html)
- Glasgow IR group Wiki DFR page (http://ir.dcs.gla.ac.uk/wiki/DivergenceFromRandomness)