Estimating Uncertainty of Categorical Web Data

Davide Ceolin, Willem Robert van Hage, Wan Fokkink, and Guus Schreiber

VU University Amsterdam
De Boelelaan 1081a
1081HV Amsterdam, The Netherlands
{d.ceolin,w.r.van.hage,w.j.fokkink,guus.schreiber}@vu.nl

Abstract. Web data often manifest high levels of uncertainty. We focus on categorical Web data and we represent these uncertainty levels as first or second order uncertainty. By means of concrete examples, we show how to quantify and handle these uncertainties using the Beta-Binomial and the Dirichlet-Multinomial models, as well as how take into account possibly unseen categories in our samples by using the Dirichlet Process.

Keywords: Uncertainty, Bayesian statistics, Non-parametric statistics, Beta-Binomial, Dirichlet-Multinomial, Dirichlet Process

1 Introduction

The World Wide Web and the Semantic Web offer access to an enormous amount of data and this is one of their major strengths. However, the uncertainty about these data is quite high, due to the multi-authoring nature of the Web itself and to its time variability: some data are accurate, some others are incomplete or inaccurate, and generally, such a reliability level is not explicitly provided.

We focus on the real distribution of these Web data, in particular of categorical Web data, regardless of whether they are provided by documents, RDF (see [27]) statements or other means. Categorical data are the among the most important types of Web data, because they include also URIs. We do not look for correlations among data, but we stick to estimating how category proportions distribute over populations of Web data.

We assume that any kind of reasoning that might produce new statements (e.g. subsumption) has already taken place. Hence, unlike for instance Fukuoe et al. (see [10]), that apply probabilistic reasoning in parallel to OWL (see [26]) reasoning, we will propose some models to address uncertainty issues on top of that kind of reasoning layers. These models, namely the parametric Beta-Binomial and Dirichlet-Multinomial, and the non-parametric Dirichlet Process, will use first and second order probabilities and the generation of new classes of observations, to derive safe conclusions on the overall populations of our data, given that we are deriving those from possibly biased samples.

First we will describe the scope of these models (section 2), second we will introduce the concept of conjugate prior (section 3), and then two classes of models: parametric (section 4) and non-parametric (section 5). Finally we will discuss the results and provide conclusions (section 6).

2 Scope of this work

2.1 Empirical evidence from the Web

Uncertainty is often an issue in case of empirical data. This is especially the case with empirical Web data, because the nature of the Web increases the relevance of this problem but also offers means to address it, as we will see in this section. The relevance of the problem is related to the utilization of the mass of data that any user can find over the network: can one safely make use of these data? Lots of data are provided on the Web by entities the reputation of which is not surely known. In addition to that, the fact that we access the Web by crawling, means that we should reduce our uncertainty progressively, as long as we increment our knowledge. Moreover, when handling our samples it is often hard to determine how representative such a sample is of the entire population, since often we do not own enough sure information about it.

On the other hand, the huge amount of Web data gives also a solution for managing this reliability issue, since it can hopefully provide the evidence necessary to limit the risk when using a certain data set.

Of course, even within the Web it can be hard to find multiple sources asserting about a given fact of interest. However, the growing dimension of the Web makes it reasonable to believe in the possibility to find more than one data set about the given focus, at least by means of implicit and indirect evidence.

This work aims showing how it is possible to address the described issues by handling such empirical data, categorical empirical data in particular, by means of the Beta-Binomial, Dirichlet-Multinomial and Dirichlet Process models.

2.2 Requirements

Our approach will need to be quite elastic in order to cover several issues, as described below. The non-triviality of the problem comes in a large part from the impossibility to directly handle the sampling process from which we derive our conclusions. The requirements that we will need to meet are:

Ability to handle incremental data acquisition The model should be incremental, in order to reflect the process of data acquisition: as long as we collect more data (even by crawling), our knowledge will reflect that increase.

Prudence It should derive prudent conclusions given all the available information. In case not enough information is available, the wide range of possible conclusions derivable will clearly make it harder to set up a decision strategy.

Cope with biased sampling The model should deal with the fact that we are not managing a supervised experiment, that is, we are not randomly sampling from the population. We are using an available data set to derive safe consequences, but these data could, in principle, be incomplete, inaccurate or biased, and we must take this into account.

Ability to handle samples from mixtures of probability distributions The data we have at our disposal may have been drawn from diverse distributions, so we can't use the central limit theorem, because it relies on the fact that the sequence of variables is identically distributed. This implies the impossibility to make use of estimators that approximate by means of the Normal distribution.

Ability to handle temporal variability of parameters Data distributions can change over time, and this variability has to be properly accounted.

Complementarity with higher order layers The aim of the approach is to quantify the intrinsic uncertainty in the data provided by the reasoning layer, and, in turn, to provide to higher order layers (time series analysis, decision strategy, trust, etc.) reliable data and/or metadata.

2.3 Related work

The models adopted here are applied in a variety of fields. For the parametric models, examples of applications are: topic identification and document clustering (see [18, 6]), quantum physics (see [15]), and combat modeling in the naval domain (see [17]). What these heterogeneous fields have in common is the presence of multiple levels of uncertainty (for more details about this, see sect. 4).

Also non-parametric models are applied in a wide variety of fields. Examples of these applications include document classification [3] and haplotype inference [30]. These heterogeneous fields have in common with the previous application the presence of several layers of uncertainty, but they also show lack of prior information about the number of parameters. These concepts will be treated in section 5 where even the Wilcoxon sign-ranked test (see [29]), used for validation purposes, falls into the non-parametric models class.

As to our knowledge, the chosen models have not been applied to categorical Web data yet. We propose to adopt them, because, as the following sections will show, they fit the requirements previously listed.

3 Prelude: Conjugate priors

To tackle the requirements described in the previous section, we adopt some bayesian parametric and non-parametric models in order to be able to answer questions about Web data.

Conjugate priors (see [12]) are the "leit motiv", common to all the models adopted here. The basic idea starts from the Bayes theorem (1): given a prior knowledge and our data, we update the knowledge into a posterior probability.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{1}$$

This theorem describes how it is possible to compute the posterior probability, $P(A|B)$, given the prior probability of our data, $P(A)$, the likelihood of the model, given the data, $P(B|A)$, and the probability of the model itself, $P(B)$.

When dealing with continuous probability distributions, the computation of the posterior distribution by means of Bayes theorem can be problematic, due to the need to possibly compute complicated integrals. Conjugate priors allow us to overcome this issue: when prior and posterior probability distributions belong to the same exponential family, the posterior probability can be obtained by updating the prior parameters with values depending on the observed sample (see also [9]). Exponential families are classes of probability distributions having their density functions sharing the form $f(x) = e^{a(q)b(x)+c(q)+d(x)}$, with q a known parameter and a, b, c, d known functions. Exponential families include many important probability distributions, like the Normal, Binomial, Beta, etc., see [5]. So, if X is a random variable that distributes as defined by the function $P(p)$ (for some parameter or vector of parameters p) and, in turn, p distributes as $Q(\alpha)$ for some parameter (or vector of parameters) α called "hyperparameter"), and P belongs to the same exponential family as Q,

$$p \sim Q(\alpha),\ X \sim P(p)$$

then, after having observed *obs*,

$$p \sim Q(\alpha')$$

where $\alpha' = f(\alpha, obs)$, for some function f. For example, the Beta distribution is the conjugate of the Binomial distribution. This means that the Beta, shaped by the prior information and by the observations, defines the range within which the parameter p of the Binomial will probably be situated, instead of directly assigning to it the most likely value. Other examples of conjugate priors are: Dirichlet, which is conjugate to the Multinomial, and Gaussian, which is conjugate to itself.

Conjugacy guarantees ease of computation, which is a desirable characteristic when dealing with very big data sets as Web data sets often are. Moreover, the model is incremental, and this makes it fit the crawling process with which Web data are obtained, because crawling, in turn, is an incremental process. Both the heterogeneity of the Web and the crawling process itself increase the uncertainty of Web data. The probabilistic determination of the parameters of the distributions adds a smoothing factor that helps to handle this uncertainty.

4 Parametric bayesian models for categorical Web data

In this section we will handle situations where the number of categories is known a priori, by using the Dirichlet-Multinomial model and its special case with two categories, i.e. the Beta-Binomial model [9]. As generalized versions of the Binomial and Multinomial distribution, they describe the realization of sequences of mutually exclusive events. Categorical data can be seen as examples of such sequences. These models are parametric, since the number and type of parameters is given a priori, and they can also be classified as "empirical bayesian models". This further classification means that they can be seen as an approximation of a full hierarchical bayesian model, where the prior hyperparameters are set to their maximum likelihood values according to the analyzed sample.

4.1 Case study 1: Deciding between alternatives - ratio estimation

Suppose that a museum has to annotate a particular item I of its collection. Suppose further, that the museum does not have expertise in the house about that particular subject and, for this reason, in order to correctly classify the item, it seeks judgments from outside people, in particular from Web users that provide evidence of owning the desired expertise.

After having collected judgements, the museum faces two possible classifications for the item, C1 and C2. C1 is supported by four experts, while C2 by only one expert. We can use these numbers to estimate a probability distribution that resembles the correct distribution of C1 and C2 among all possible annotations.

A basic decision strategy that could make use of this probability distribution, could accept a certain classification only if its probability is greater or equal to a given threshold (e.g. 0.75). If so, the Binomial distribution representing the sample would be treated as representative of the population, and the sample proportions would be used as parameters of a Bernoulli distribution about the possible classifications for the analyzed item: $P(class(I) = C1) = 4/5 = 0.8$, $P(class(I) = C2) = 1/5 = 0.2$. (A Bernoulli distribution describes the possibility that one of two alternative events happens. One of these events happens with probability p, the other one with probability $1 - p$. A Binomial distribution with parameters n, p represents the outcome of a sequence of n Bernoulli trials having all the same parameter p.)

However, this solution shows a manifest leak. It provides to the decision strategy layer the probabilities for each of the possible outcomes, but these probabilities are based on the current available sample, with the assumption that it correctly represents the complete population of all existing annotations. This assumption is too ambitious. (Flipping a coin twice, obtaining a heads and a tails, does not guarantee that the coin is fair, yet.)

In order to overcome such a limitation, we should try to quantify how much we can rely on the computed probability. In other words, if the previously computed probability can be referred as a "first order" probability, what we need to compute now is a "second order" probability (see [15]). Given that the conjugate prior for the Binomial distribution representing our data is the Beta distribution, the model becomes:

$$p \sim Beta(\alpha, \beta),\ X \sim Bin(p, n) \tag{2}$$

where $\alpha = \#evidence_{C1} + 1$ and $\beta = \#evidence_{C2} + 1$.

By analyzing the shape of the conjugate prior Beta(5,2), we can be certain enough about the probability of C1 being safely above our acceptance threshold. In principle, our sample could be drawn by a population distributed with a $40\% - 60\%$ proportion. If so, given the threshold of acceptance of 0.75, we would not be able to take a decision based on the evidence. However, the quantification of that proportion would only be possible if we know the population. Given that we do not have such information, we need to estimate it, by computing (3), where we can see how the probability of the parameter p being above the threshold is less than 0.5. This manifests the need for more evidence: our sample suggests to accept the most popular value, but the sample itself does not guarantee to be representative enough of the population.

$$P(p \geq 0.75) = 0.4660645,\ \ p \sim Beta(5,2) \tag{3}$$

Table 1 shows how the confidence in the value p being above the threshold grows as long as we increase the size of the sample, when the proportion is kept. By applying the previous strategy (0.75 threshold) also to the second order probability, we will still choose C1, but only if supported by a sample of size at least equal to 15.

Table 1: The proportion within the sample is kept, so the most likely value for p is always exactly that ratio. However, given our 0.75 threshold, we are sure enough only if the sample size is 15 or higher.

#C1	$\#C2$	$P(p \geq 0.75)$
		$p \sim Beta(\#C1+1, \#C2+1)$
4	1	0.4660645
8	2	0.5447991
12	3	0.8822048

Finally, these considerations could also be done on the basis of the Beta-Binomial distribution, which is a probability distribution representing a Binomial which parameter p is randomly drawn from a Beta distribution. The Beta-Binomial summarizes model (2) in one single function (4). We can see from Table 2 that the expected proportion of the probability distribution approaches the ratio of the sample (0.8), as the sample size grows. If so, the sample is regarded as a better representative of the entire population and the Beta-Binomial, as sample size grows, will converge to the Binomial representing the sample (see Fig. 1).

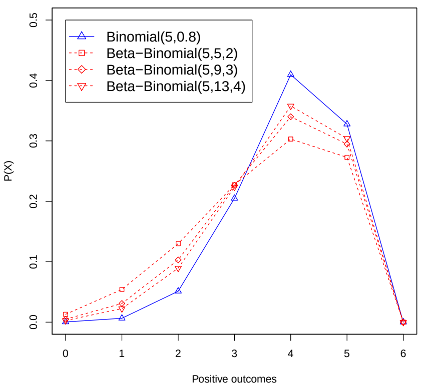


Fig. 1: Comparison between Binomial and Beta-Binomial with increasing sample size. As the sample size grows, Beta-Binomial approaches Binomial.

$$X \sim BetaBin(n, \alpha, \beta) = p \sim Beta(\alpha, \beta), X \sim Bin(n, p) \tag{4}$$

4.2 Case study 2: deciding proportions - confidence intervals estimation

The Linked Open Piracy[1] is a repository of piracy attacks that happened around the world in the period 2005 - 2011, derived from reports retrieved from the ICC-

[1] http://semanticweb.cs.vu.nl/lop

X	$E(X)$	$p = E(X)/n$
BetaBin(5,5,2)	3.57	0.71
BetaBin(5,9,3)	3.75	0.75
BetaBin(5,13,4)	3.86	0.77

Table 2: The sample proportion is kept, but the "expected proportion" p of Beta-Binomial passes the threshold only with a large enough sample. $E(X)$ is the expected value.

CCS website.[2] Attack descriptions are provided, in particular covering their type (boarding, hijacking, etc.), place, time, as well as ship type.

Data about attacks is provided in RDF format, and a SPARQL (see [28]) endpoint permits to query the repository. Such a database is very useful, for instance, for insurance companies to properly insure ships. The premium should be related to both ship conditions and their usual route. The Linked Open Piracy repository allows an insurance company to estimate the probability to be victim of a particular type of attacks, given the programmed route. Different attack types will imply different risk levels.

However, directly estimating the probability of a new attack given the dataset, would not be correct, because, although derived from data published from an official entity like the Chamber of Commerce, the reports are known to be incomplete. This fact clearly affects the computed proportions, especially because it is likely that this incompleteness is not fully random. There are particular reasons why particular attack types or attacks happening in particular zones are not reported. Therefore, beyond the uncertainty about the type of next attack happening (first order uncertainty), there will be an additional uncertainty order due to the uncertainty in the proportions themselves. This can be handled by a parametric model that will allow to estimate the parameters of a Multinomial distribution. The model that we are going to adopt is the multivariate version of the model described in section 4, that is, the Dirichlet-Multinomial model (see [6, 17, 18]):

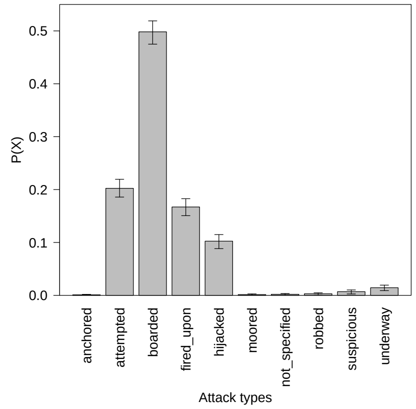


Fig. 2: Attack type proportion and confidence intervals

$$Attacks \sim \text{Multinom}(params),\ params \sim \text{Dirichlet}(\alpha) \tag{5}$$

where α is the vector of observations per attack type (incremented by one unit each, as the α and β parameters of Beta probability distribution). By adopting this model, we are able to properly handle the uncertainty carried by our sample, due to either time variability (over the years, attack type proportions could have changed) or biased samples. Drawing the parameters of our Multinomial

[2] http://www.icc-ccs.org/

distribution from a Dirichlet distribution instead of directly estimating them, allows us to compensate for this fact, by smoothing our attacks distribution. As a result of the application of this model, we can obtain an estimate of confidence intervals for the proportions of the attack types (with 95% of significance level, see (6)). These confidence intervals depend both on the sample distribution and on its dimension (Fig. 2).

$$\forall p \in param, CI_p = (p - \theta_1, p + \theta_2), P(p - \theta_1 \leq p \leq p + \theta_2) = 0.95 \qquad (6)$$

5 Non-parametric bayesian models

In some situations, the previously described parametric models do not fit our needs, because they set a priori the number of categories, but this is not always possible. In the previous example, we considered and handled uncertainty due to the possible bias of our sample. The proportions showed by our sample could be barely representative of the entire population because of a non-random bias, and therefore we were prudent in estimating densities, even not discarding entirely those proportions. However, such an approach lacks in considering another type of uncertainty: we could not have seen all the possible categories and we are not allowed to know all of them a priori. Our approach was to look for the prior probability to our data in the n-dimensional simplex, where n is the number of categories, that is, possible attack types. Now such an approach is no more sufficient to address our problem. What we should do is to add yet another hierarchical level and look for the right prior Dirichlet distribution in the space of the probability distributions over probability distributions (or space of simplexes). Non-parametric models differ from parametric models in that the model structure is not specified a priori but is instead determined from data. The term non-parametric is not meant to imply that such models completely lack parameters, but that the number and nature of the parameters are flexible and not set in advance. Hence, these models are also called "distribution free".

5.1 Dirichlet Process

Dirichlet Processes [8] are a generalization of Dirichlet distributions, since they correspond to probability distributions of Dirichlet probability distributions. They are stochastic processes, that is, sequences of random variables (distributed as Dirichlet distributions) which value depends on the previously seen ones. Using the so-called "Chinese Restaurant Process" representation (see [22]), it can be described as follows:

$$X_n = \begin{cases} X_k^* & \text{with probability } \frac{num_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } H & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases} \qquad (7)$$

where H is the continuous probability measure ("base distribution") from which new values are drawn, representing our prior best guess. Each draw from H will return a different value with probability 1. α is an aggregation parameter, inverse of the variance: the higher α, the smaller the variance, which can be interpreted as the confidence value in the base distribution H: the higher the α value is, the more the Dirichlet Process resembles H. The lower the α is, the more the value of the Dirichlet Process will tend to the value of the empirical distribution observed. Each realization of the process is discrete and is equivalent to a draw from a Dirichlet distribution, because, if

$$G \sim DP(H, \alpha) \tag{8}$$

is a Dirichlet Process, and $\{B\}_{i=1}^{n}$ are partitions of S, we have that

$$(G(B_1)...G(B_n)) \sim Dirichlet(\alpha H(B_1)...\alpha H(B_n)) \tag{9}$$

If our prior Dirichlet Process is (8), given (9) and the conjugacy between Dirichlet and Multinomial distribution, our posterior Dirichlet Process (after having observed n values θ_i) can be represented as one of the following two representations:

$$(G(B_1)...G(B_n))|\theta_1...\theta_n \sim Dirichlet(\alpha H(B_1) + n_{\theta_1}...\alpha H(B_n) + n_{\theta_n}) \tag{10}$$

$$G \mid \theta_1...\theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\Sigma_{i=1}^{n} \delta_{\theta_i}}{n}\right) \tag{11}$$

where δ_{θ_i} is the Dirac delta function (see [4]), that is, the function having density only in θ_i. The new base function will therefore be a merge of the prior H and the empirical distribution, represented by means of a sum of Dirac delta's. The initial status of a Dirichlet Process posterior to n observations, is equivalent to the *nth* status of the initial Dirichlet Process that produced those observations (see De Finetti theorem, [13]).

The Dirichlet process, starting from a (possibly non-informative) “best guess”, as long as we collect more data, will approximate the real probability distribution. Hence, it will correctly represent the population in a prudent (smoothed) way, exploiting conjugacy like the Dirichlet-Multinomial model, that approximates well the real Multinomial distribution only with a large enough data set (see section 4). The improvement of the posterior base distribution is testified by the increase of the α parameter, proportional to the number of observations.

5.2 Case study 3: Classification of piracy attacks - unseen types generation

We aim at predicting the type distributions of incoming attack events. In order to build an “infinite category” model, we need to allow for event types to be randomly drawn from an infinite domain. Therefore, we map already observed attack types with random numbers in [0..1] and, since all events are a priori equally likely, then new events will be drawn from the Uniform distribution, $U(0,1)$, that is our base distribution (and is a measure over [0..1]). The model then is:

- $type_1 \sim DP(U(0,1), \alpha)$: the prior over the first attack type in region R;
- $attack_1 \sim Categorical(type_1)$: type of the first attack in R during $year_y$.

After having observed $attack_{1 \ldots n}$ during $year_y$, our posterior process becomes:

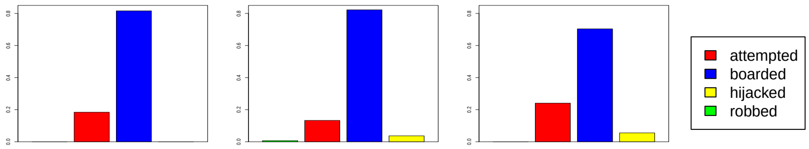
$$type_{n+1} \mid attack_{1 \ldots n} \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} U(0,1) + \frac{n}{\alpha + n} \frac{\Sigma_{i=1}^{n} \delta_{attack_i}}{n}\right)$$

where α is a low value, given the low confidence in $U(0,1)$, and $type_{n+1}$ is the prior of $attack_{n+1}$, that happens during $year_{y+1}$. A Categorical distribution is a Bernoulli distribution with more than two possible outcomes (see Section 4).

Results Focusing on each region at time, we simulate all the attacks that happened there in $year_{y+1}$. Names of new types generated by simulation are matched to the actual $year_{y+1}$ names, that do not occur in $year_y$, in order of decreasing probability. The simulation is compared with a projection of the proportions of $year_n$ over the actual categories of $year_{n+1}$. The comparison is made by measuring the distance of our simulation and of the projection from the real attack types proportions of $year_{y+1}$ using the the Manhattan distance (see [16]). This metric simply sums, for each attack type, the difference between the real $year_{y+1}$ probability and the one we forecast. Hence, it can be regarded as an error measure. Table 3 summarizes the results over the entire dataset.[3] Our simulation reduces the distance (i.e. the error) with respect to the projection, as confirmed by a Wilcoxon signed-rank test [29] at 95% significance level. (This non-parametric statistical hypothesis test is used to determine whether one of the means of the population of two samples is smaller/greater than the other.) The simulation improves when large amount of data is available and the category cardinality varies, as in case of Region India, which results are reported in Fig. 3 and 4a.

Table 3: Averages and variances of the error of the two forecasts. The simulation gets a better performance.

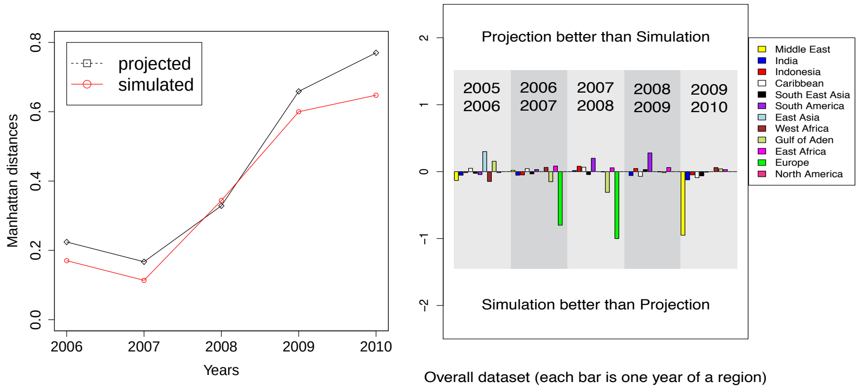
	Simulation	Projection
Average distance	0.29 △	0.35
Variance	0.09 △	0.21



(a) Year 2005 (b) Simulation (c) Year 2006

Fig. 3: Comparison between the projection forecast and the simulation forecast with the real-life year 2006 data of region India.

[3] The code can be retrieved at `http://www.few.vu.nl/~dceolin/DP/Dir.R`



(a) Errors

(b) Distances differences

Fig. 4: Error distance from real distribution of the region India (fig. 4a) and differences of the error of forecast based on simulation and on projection (fig. 4b). Positive difference means that the projection predicts better than our simulation.

6 Conclusions and future work

The fact that our proposed models fit well with the expressed requirements is apparently a good hypothesis to continue to explore, because we have seen how it is possible to handle such uncertainty and to transform it in a smoothing factor of the probability distribution that we estimate given our evidence, by allowing the parameters of our distributions to be probabilistically determined. Moreover, we have built models able to produce reliable forecasts also when not every class is know a priori. We also provided case study validation of the suggested models.

The set of models will be extended to deal with concrete domain data (e.g. time intervals, measurements), for instance, by adopting the Normal or the Poisson Process (see [9]). Moreover, automatic model selection will be investigated, in order to choose the best model also when the limited information about our problems could make more models suitable. From a pure Web perspective, our models will be extended to properly handle contributions coming from different sources together with their reputation. This means, on one side, considering also provenance (like in [1]) and, on the other side, using Mixture Models ([23]), Nested ([24]) and Hierarchical Dirichlet Processes ([25]), eventually employing Markov Chain Monte Carlo algorithms (see [7, 21]) to handle lack of conjugacy.

References

1. D. Ceolin, P. Groth, and W. R. van Hage. Calculating the trust of event descriptions using provenance. In *SWPM 2010 Proceedings*, 2010.
2. D. Ceolin, W.R. van Hage, and W. Fokkink. A trust model to estimate the quality of annotations using the web. In *WebSci10*, 2010.

3. M. Davy and J. Tourneret. Generative supervised classification using dirichlet process priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1781–1794, 2010.
4. P. Dirac. *Principles of quantum mechanics.* Oxford at the Clarendon Press, 1958.
5. Andersen E. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65:1248–1255, 9 1970.
6. C. Elkan. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *ICML*, volume 148, pages 289–296. ACM, 2006.
7. M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1994.
8. T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.
9. D. Fink. A Compendium of Conjugate Priors. Technical report, Cornell University, 1995.
10. A. Fokoue, M. Srivatsa, and R. Young. Assessing trust in uncertain information. In *ISWC*, pages 209–224, 2010.
11. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* CRC Press, 2003.
12. R. Schlaifer H. Raiffa. *Applied statistical decision theory.* M.I.T. Press, 1968.
13. M. Hazewinkel. *Encyclopaedia of Mathematics*, chapter De Finetti theorem. Springer, 2001.
14. B. He, M. Patel, Z. Zhang, and K. C. Chang. Accessing the deep web. *Commun. ACM*, 50:94–101, May 2007.
15. J. Hilgevoord and J. Uffink. Uncertainty in prediction and in inference. *Foundations of Physics*, 21:323–341, 1991.
16. E. F. Krause. *Taxicab Geometry.* Dover, 1987.
17. P. Kvam and D. Day. The multivariate polya distribution in combat modeling. *Naval Research Logistics (NRL)*, 48(1):1–17, 2001.
18. R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In *ICML*, ICML '05, pages 545–552. ACM, 2005.
19. T. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft Research, 2003.
20. P. Müller N. L. Hjort, C. Holmes and S. G. Walker. *Bayesian Nonparametrics.* Cambridge University Press, 2010.
21. R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and graphical statistics*, 9(2):249–265, 2000.
22. J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102(2):145–158, 1995.
23. Carl Edward Rasmussen. The infinite gaussian mixture model. In *In Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.
24. A. Rodriguez, D. B. Dunson, and A. E. Gelfand. The nested dirichlet process. *Journal of the American Statistical Assoc.*, 103(483):1131–1144, September 2008.
25. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Assoc.*, 101(476):1566–1581, 2006.
26. W3C. OWL Reference, August 2011. http://www.w3.org/TR/owl-ref/.
27. W3C. Resource Definition Framework, August 2011. http://www.w3.org/RDF/.
28. W3C. SPARQL, August 2011. http://www.w3.org/TR/rdf-sparql-query/.
29. F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.
30. E. Xing. Bayesian Haplotype Inference via the Dirichlet Process. In *ICML*, pages 879–886. ACM Press, 2004.