Topic Models and Advanced Algorithms for Profiling of Knowledge in Scientific Papers

V. Jelisavčić*, B. Furlan **, J. Protić**, V. Milutinović**
* Mathematical Institute of the Serbian Academy of Sciences and Arts/11000, Belgrade, Serbia
** Department of Computer Engineering, School of Electrical Engineering,
University of Belgrade/11000, Belgrade, Serbia

vladisavj@gmail.com, {bojan.furlan, jeca, vm}@etf.rs

Abstract - Survey of probabilistc topic models is presented with emphasis on fundamentally different approaches used in modeling. Introduced classification differs from earlier efforts, providing a complementary view of the field. Purpose of this survey is to provide a brief overview of the current probailistic topic models as well as an inspiration for future research.

I. INTRODUCTION

Probabilistic topic models are a group of machine learning algorithms for discovering latent topical structures in data. Although many applications are found in various data mining areas such as image annotation, audio and video analysis, they are primarily invented for use in finding topics in textual data. Profiling and modeling knowledge from scientific papers is one area of research that benefits most.

Few surveys of topic models already exist; among most significant are [1], [2] and [3]. Most notable one [1], presents a classificaion of directed probabilistic topic models and a broader view on graphical models in general, and can serve as a great starting point for venture in the field of topic modeling. In [2] and [3] a gentle introduction is made to the field.

Main criterion of classification in [1] is functionality, and models are presented in a chronological order in a systematic evolution-based fashion. This is not the purpose of this survey; functionality is not of a primary interest for us. Criteria of presented classification are chosen as to highlight fundamental approaches and assumptions used in topic modeling. Also, [1] focuses on directed probabilistic topic models while we impose no such restriction. Introducing general ideas and formal definition has also been done in [2] and [3] so this is not our primary goal either. In [1] models are also classified acording to their original problem domain. As many of those problems, such as topic discovery, topic evolution, document classification and many others, present a subproblem to the modeling and profiling of the knowledge from scientific papers, such distinction is not made in this survey.

Survey of topic models is presented with emphasis on different approaches used.

II. CLASSIFICATION

Topic models are classified according to three orthogonal critera. First criterion is based on word ordering and a document representation. Two distinct approaches are posible. Simpler and very often more usefull solution is commonly known as bag of words. In this document representation the word ordering is neglected which enables focus on a global semantic structures without need to model local word order dependencies. Other approach, that doesn't neglect word ordering will be referred to as a sequence of words. Although first approach is appreciated for its simplicity and is often sufficient, second approach bear more information which can supposedly lead to better results in some problem domains. Second criterion is taking external knowledge into consideration. First approach where no such knowledge is provided is simpler and for many purposes sufficient. Second approach is based on using in-domain knowledge for the target problem, yielding more specific and human interpretable topics. Third and final criterion is dependability on labeled data. Main idea behind topic models is unsupervised clustering of topics which renders them applicable to a broad range of real life problems where there are no data labels and cannot be provided. Most of the topic models are fully unsupervised. Some models can be used in a supervised or semi-supervised manner in order to be applicable to classification tasks or simply to yield better results if labeled data for domain is already present. Classification tree with corresponding models at the leaves can be seen on Figure 1.

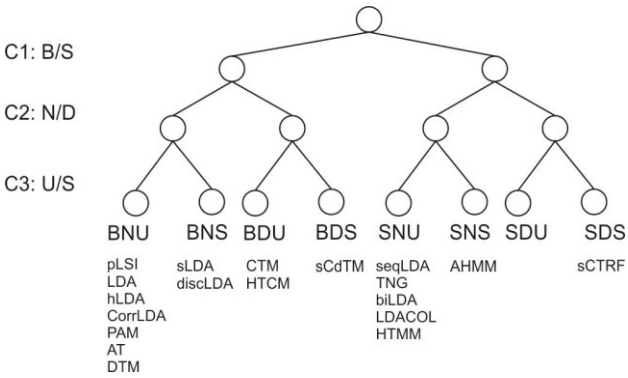

Figure 1. The classification three of probabilistic topic models. **Legend**: B/S – bag of words vs sequence of words; N/D – no in-domain vs in-domain; U/S – unsupervised vs supervised. **Description:** The classification three obtained by successive application of the chosen criteria. **Implication:** The class of unsupervised sequence of word models with in-domain knowledge requirements have no known implementations.

	C1: B/S	C2: N/D	C3: U/S	Ability
pLSI [Hoffman 1999]	B	N	U	
LDA [Blei 2003]	B	N	U	
hLDA [Blei 2003]	B	N	U	topic hierarchy, number of topics not fixed
DTM [Blei 2006]	B	N	U	time evolution of topics
CorrLDA [Blei 2006]	B	N	U	topic correlations as matrix
PAM [Li 2006]	B	N	U	topic correlations as DAG
ATM [Zvi 2010]	B	N	U	topic authorship
sLDA [Blei 2007]	B	N	S	supervised learning of topics
DMR [Mimno 2008]	B	N	S	arbitrary document metadata
CTM [Steyvers 2011]	B	D	U	arbitrary word level features
sCdTM [Zhu 2010]	B	D	S	arbitrary word level features, supervised
TNG [Wang 2007]	S	N	U	phrases and n-grams
AHMM [Blei 2001]	S	N	S	
sCTRf [Zhu 2010]	S	D	S	arbitrary word level features, supervised, sequential
discLDA [Xu 2010]	B	N	S	supervised learning of topics
HTCM [Steyvers 2011]	B	D	U	arbitrary word level features, hierarchy of features
biLDA [Wallach 2006]	S	N	U	bigrams
LDACOL []	S	N	U	unigrams and bigrams
SeqLDA [Du 2010]	S	N	U	

III. EXISTING SOLUTIONS

For each class defined in previous section most prominent examples are presented, if such solutions exist.

A. Unsupervised bag of words topic models with no in-domain knowledge requirements

This class of models reside on word exchangeability assumption, i.e. discards information on word position within documents. Such models are often used regarding problems such as information retrieval, document clustering and summarization due to their simplicity introduced with bag of words approach and greater real-world problem applicability based on their unsupervised nature. Most of the probabilistic topic models, including the earliest ones, fall into this category. There are numerous extensions from the baseline approach (Latent Dirichlet Allocation) that introduce additional abilities beside modeling word-topic and document-topic distributions, some of which are presented here.

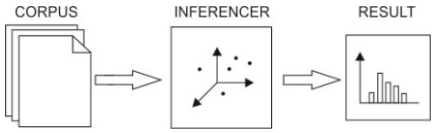

Figure 2. Outline of the unsupervised bag of words topic model with no in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Word ordering is neglected.

A.1. Probabilistic Latent Semantic Indexing

Probabilistic Latent Semantic Indexing (pLSI) was invented by T. Hoffman and was first published in Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval in 1999 as a probabilistic variant of Latent Semantic Analysis that has a sound statistical foundation and defines a proper generative data model [15].

pLSI models generative process responsible for creating each document in corpus, where each word in a document is sampled from a mixture of multinomial distributions that can be interpreted as topics, and proportions corresponding to mixture weights are sampled from a separate multinomial distribution for each document. Based on generative model, an inference algorithm is defined as a method for infering topic-word distributions, as well as document-topic distributions, from textual corpora.

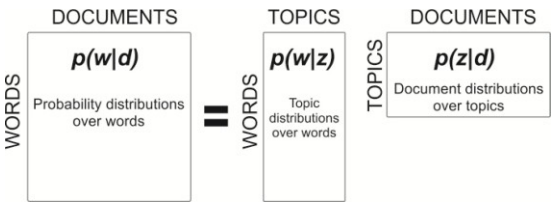

Figure 3. Probabilistic Latent Semantic Indexing, viewed as a matrix factorization

There are several methods for computing word-topic and topic-document distributions, one videly accepted is Expectation Maximization algorithm. Equations for E and M steps are inferred directly from the generative model (Fig. 4).

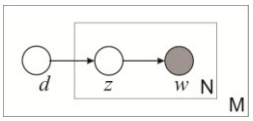

Figure 4. Plate notation of pLSI model. **Description**: Graphical presentation of a bayesian network corresponding to pLSI model. For interpretation of this as well as other graphical models presented in the survey, reader is encouraged to read [1] and [2].

pLSI efficiently resolve several issues of Latent Semantic Analysis (LSA) [16], it's non-probabilistic predecessor, such as capturing polysemy. Also, as opposed to LSA, this generative model has a strong theoretical justification. Problem that pLSI is often confronted to is large number of estimation parameters that depends on corpus size which can create problems with overfiting as number of documents increases, as well as inability to be applied incrementally to unseen documents due to its offline nature.

A.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that represents a bayesian upgrade to pLSI by introducing priors on document-topic distributions and is published in Journal of Machine Learning Research by D. Blei, A. Ng, and M. Jordan in 2003 [9] creating a foundation for numerous latent structure discovery algorithms collectively known as Probabilistic topic models.

LDA resolves problematic issues of pLSI such as increasing number of estimation parameters by placing a Dirichlet prior distribution, effectively considering them as a random variables on their own. By introducing such a prior not only the number of estimation parameters was reduced (and made independant of the number of documents), but inability of model to be applied incrementally to unseen documents was also surpassed.

Because of it's increased complexity in comparison to pLSI, exact inference is intractable from the generative model (Figure 5. **LDA model in plate notation**). To efficiently cope with this problem, several approximate inference algorithms are derived such as Variational Inference, and various Markov Chain Monte Carlo algorithms, such as Gibbs Sampling[8].

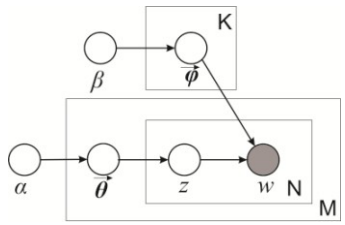

Figure 5. LDA model in plate notation

One of main advantages of LDA over earlier probabilistic methods such as pLSI, beside ability to be incrementally applied to unseen documents, i.e. online property, lies in its expandability. LDA serves as a basis for many topic models, some of which are presented in further text. LDA is computationally more expensive than earlier models such as pLSI and LSA, but also lacks some of the features of later more complex models such as modeling relationships between topics [13][14][23], modeling evolution of topics over time based on document metadata [11][12], modeling autorship [19], modeling arbitrary document metadata [18] and others. As a attempt to address the computational time requirements, several implementations exploiting potentional parallelism are made [22].

A.3. Hierarchical Latent Dirichlet Allocation

Hierarchical LDA is introduced by Blei et al in 2003 as a extension to LDA that can model a tree of topics instead of a flat topic structure introduced by LDA [13].

Hierarchical LDA uses non-parametric bayesian approach to model topical hierarchies. Tree of topics is defined procedurally by an algorithm that constructs hierarchy as data are made available. Every node in the topic tree represents a random variable, and each has a word-topic distribution assigned. Document can be generated by traversing the tree from root to one of it's leaves while sampling topics along the path.

Graphical model for hLDA can be seen in Figure 6.

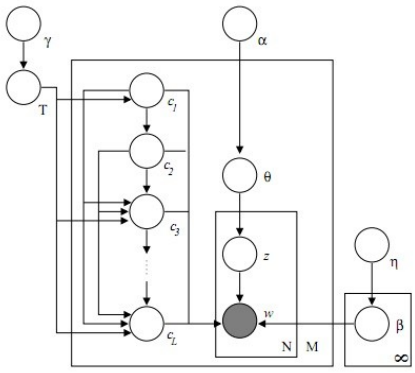

Figure 6. Hierarchical Latent Dirichlet Distribution

A.4. Dynamical Topic Model

Dynamical Topic Model (DTM) are introduced by D. Blei and J. Laferty in Proceedings of the 23rd international conference on Machine learning in New York, USA 2006., as an enhancement to Latent Dirichlet Allocation which enabled modeling of topic evolution in time [12].

Dynamical topic model includes notion of time in topic modeling using document metadata and therefore can discribe evolution of word-topic distributions. Using this approach topic trends can be observed.

As a extension to LDA, Dynamical topic model (Figure 7) yields more complicated inference, and because of non-conjugacy, sampling methods are more dificult to infere, so variational methods such as Variational Kalman Filtering or Variational Wavelet Regression are used [12].

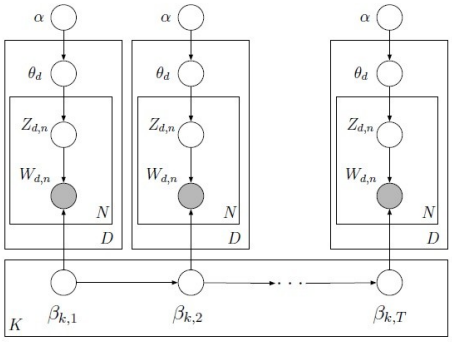

Figure 7. Dynamical Topic Model in plate notation

Advantage of DTM is ability to track topics through time, which was impossible using previous probabilistic topic modeling algorithms. Some of most significant disadvantages of Dynamical Topic Model are fixed number of topics and a discrete notion of time. In many corpora topics are born and extingushed which is a behavior not properly modeled by DTM because of fixed number of topics. Complexity of variational inference for DTM grows quickly with increase in time granularity which poses a problem in determing an apropriate resolution because of memory and computational requirements

A.5. Correlated topic Model

Correlated Topic Model (CorrLDA) is an probabilistic topic model that enhances base LDA with modeling of correlations between topics, and is introduced by D. Blei and J. Laferty in Processing Systems, Proceedings of the 2005 Neural Information Processing Systems NIPS [14].

CorrLDA can model complex structure of underlying topics in textual corprora, and provide a graph representation of topic relationships, as opposed to LDA model that imposes a strong mutual independance assumption on topics, thus making it more expressive.

Generative model is represented in plate notation in Figure 8, upon which appropriate mean-field variational inference algorithm can be based [14].

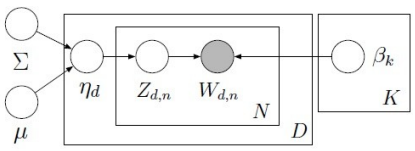

Figure 8. Correlated Topic Model, plate notation

CorrLDA provides more expressivity than LDA approach and generaly can make a better fit to some textual corpora due to more assumptions made. By modeling relations between topics, CorrLDA provides an effective way for topic visualization and exploration.

A.6. Pachinko Allocation Model

Pachinko Allocation Model was first introduced by Wei Li and Andrew McCallum in 2006. in Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh as a flexible alternative to Correlated Topic Model [23].

Like Correlated Topic Model, PAM can model correlations between topics. As opposed to CorrLDA where topic correlations are modeled using covariance matrix representing pairwise correlations between topics, PAM redefines the concept of topic as a distribution not only over words, but as a distribution over words and other topics also. This approach enables modeling arbitrary DAG topic structure that cannot be modeled using CorrLDA.

Generative model is represented in plate notation in Figure 9, and appropriate inference can be done using Gibbs Sampler.

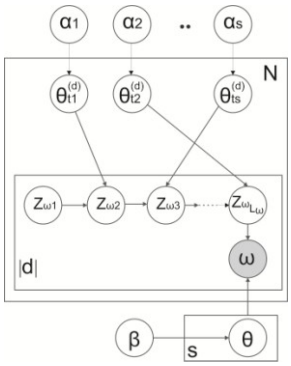

Figure 9. Plate notation of Pachinko Allocation Model

Using different approach but with the same objective, Pachinko Allocation Model provides several benefits over Correlated Topic Model. PAM can capture nested and n-ary correlations and the choice of underlying distribution is not restricted to logistic normal distribution. Also, CorrLDA must estimate parameters for each pair of topics so number of parameters grows as the square of the number of topics whereas PAM successfully avoids this problem.

A.7. Author topic Model

Author Topic Model (ATM) is a generative probabilistic topic model introduced by M. R. Zvi et al in ACM Trans. Inf. Syst., in 2010. derived from LDA as a model for detecting topics distribution corresponding to each author in textual corpora, based on metadata [19].

ATM is dependant on metadata associated with each document in corpus. Instead of modeling only document-topic and topic-word distributions, ATM goes step further and models author-topic distributions.

Author-topic and topic-word distributions can be learned using Markov Chain Monte Carlo algorithm based on generative model Figure **10**10.

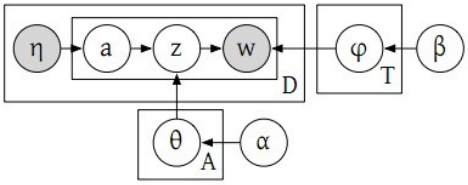

Figure 10. Plate notation of Author Topic Model

B. Supervised bag of words models with no in-domain knowledge requirements

This class of models stems from unsupervised bag of words models with no in-domain knowledge requirements, as a group of models used for classification instead of clustering. Due to their supervised nature, on some tasks these models can exibit better modeling results.

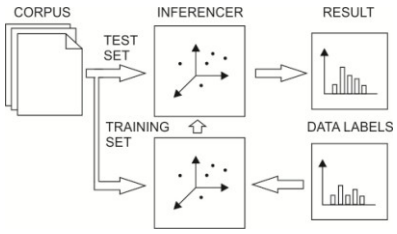

Figure 11. Outline of the supervised bag of words topic model with no in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the particular model are applied to training set given appropriate set of data labels effectively learning latent parameters. Finally, model with inferred parameters can be used for evaluation on test set or completly new set of unknown, unlabeled data. Word ordering is neglected.

B.1. Supervised LDA

Supervised Latent Dirichlet Allocation is first introduced by Blei and McAuliffe in 2007 as a supervised extension to LDA [17].

As opposed to other probabilistic topic models that work in purely unupervised fashion, sLDA extends on LDA by introducing a observable response variable in the model for each document. This extension enables sLDA to fit latent topics that will best predict future unlabeled documents.

Most appropriate aproximate inference method used for estimating the uknown parameters is Mean Field variational inference and can be derived from graphical model in Figure 127.

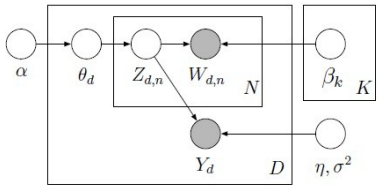

Figure 12. Supervised Topic Model

B.2. Dirichlet Multinomial Regression

Dirichlet Multinomial Regression is presented by Mimno and McCallum in a paper "Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression," in Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI '08), 2008 [18] as a extension to LDA that can incorporate various document metadata.

As opposed to previous probabilistic topic models that account for document metadata, DMR is able to incorporate arbitrary types of document metadata without additional coding. This is achieved by conditioning on metadata, rather than generating metadata or estimating metadata topical densities.

Gibbs sampler for this model can be derived based on graphical model in Figure 13.

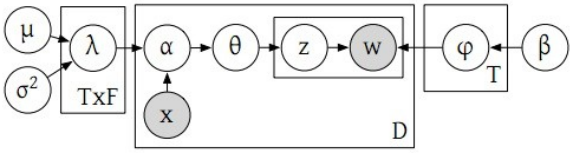

Figure 13. Multinomial Dirichlet Regression

C. Unsupervised bag of words models with in-domain knowledge requirements

Models that belong to this category make abundant use of in-domain knowledge while retaining unsupervised learning strategy. This approach is used to increase the human interpretability of topics. For instance, if modeling of a biology corpus is required, additional constraints induced by a biological ontology are expected to yield better resutls.

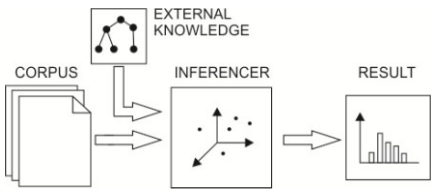

Figure 14. Outline of the unsupervised bag of words topic model with in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Additional in-domain knowledge is supplied, usually in form of ontolgy or thesaurus. Word ordering is neglected.

C.1. Concept Topic Model

Concept Topic Model was conceived by M. Steyvers et al. in 2009. as an attempt to introduce semantically rich concepts into the probabilistic model.

CTM is an extension to LDA where beside ordinary learned topics also exists a number of constrained topics where non-zero probabilities can be assigned only to

words representing human defined concepts that are provided along textual data.

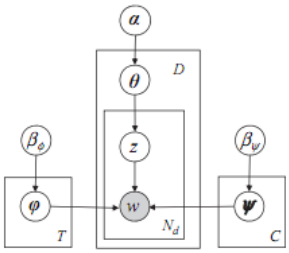

Figure 9. Concept Topic Model

D. Supervised bag of words models with in-domain knowledge requirements

This group of topic models attempt to employ additional constraints from domain of interest in classification tasks, while retaining simplicity of the bag of words assumption.

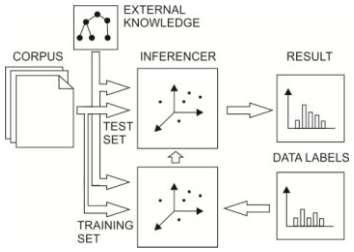

Figure 15. Outline of the supervised bag of words topic model with in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the particular model are applied to training set given appropriate set of data labels effectively learning latent parameters. Additional in-domain knowledge is supplied, usually in form of ontolgy or thesaurus. Finally, model with inferred parameters can be used for evaluation on test set or completly new set of unknown, unlabeled data. Word ordering is neglected.

D.1. *Supervised Conditional Topic Model*

Supervised Conditional Topic Model (sCdTM) is proposed by J.Xu and E.Xing in 2010. as an attempt to utilize nontrivial input features in order to improve performance.

As opposed to Dirichlet Multinomial Regression [18], that can utilize arbitrary document-level metadata, Supervised Conditional Topic Model can utilze metadata at word level which enables use of ritch feature such as POS taggs and ontologies in modeling. This is accomplished through conditioning on metadata insted of a generative approach.

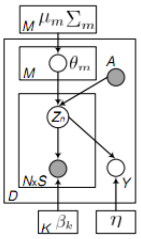

Figure 16. Conditional Topic Model in plate notation

E. Unsupervised sequence of words models with no in-domain knowledge requirements

Models belonging to this group go beyond bag of words model and account for sequential nature of textual data. Unsupervised nature of these models make them applicable to many real-world problems where data labels aren't at disposal. Lack of in-domain knowledge requirements makes them simpler and more appicable to some problems with presumably less domain-specific and humanly intepretable results.

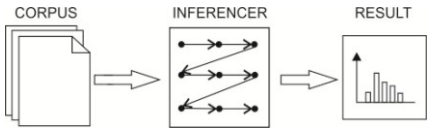

Figure 17. Outline of the unsupervised sequence of words topic model with no in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Word order is not neglected.

E.1. Topical N-Grams

Topical N-Grams (TNG) is defined by X. Wang et al in 2007. and published in Proceeding of the seventh IEEE International Conference On Data Mining, as a generative probabilistic model that attempts to relieve bag of words assumption made by Latent Dirichlet Allocation [10].

As opposed to LDA, which relies on bag of words assumption and models only unigrams, TNG also models Ngrams up to arbitrary N. Using this approach, although still relying on bag of words assumption, TNG attempts to account for sequential nature of text and enable modeling of complex phrases as well as unigrams.

Inference is slightly more complicated than in LDA, but similar aproximate inference algorithms are still applicable. Structure of TNG generative model is given in Figure 18.

Benefits of Topical NGrams model are semantically richer topic representations, enabling modeling of concepts made of multiple words which was impossible by earlier probabilistic topic models, but such benefits come at a greater computational costs.

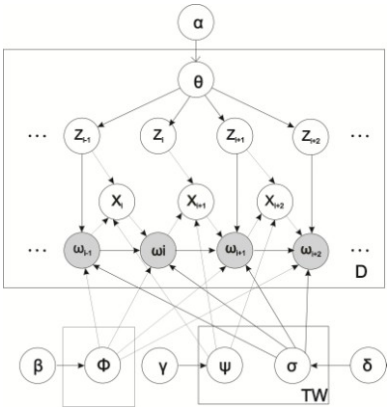

Figure 18. Plate notation of Topical NGrams model

F. Supervised sequence of words models with no in-domain knowledge requirements

Analog to supervised variants of bag of words models, these models are intended for use in classification tasks, i.e. tasks where labels corresponding to training data are provided. These models pose no in-domain knowledge requirements.

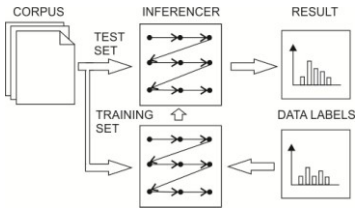

Figure 19. Outline of the supervised sequence of words topic model with no in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the particular model are applied to training set given appropriate set of data labels effectively learning latent parameters. Finally, model with inferred parameters can be used for evaluation on test set or completly new set of unknown, unlabeled data. Word order is not neglected.

F.1. Aspect Hidden Markov Model

Aspect Hidden Markov Model (AHMM) is invented by D.Blei and P. Moreno in 2001. as an attempt to use Hidden Markov Models for topic modeling.

AHMM is based on segmenting Hidden Markov Model and providing intuitive topical dependency between words and cohesive segmentation model.

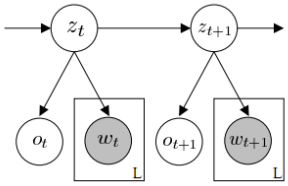

Figure 20. Plate notation of Aspect Hidden Markov Model

G. Supervised sequence of words models with in-domain knowledge requirements

This category of models account for sequential nature of textual data in supervised manner, using preassigned labels for training set while seeking to increase result specificality for domain of interest using some sort of in-domain knowledge .

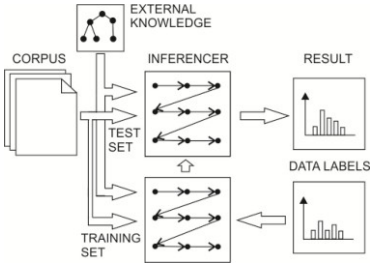

Figure 21. Outline of the supervised sequence of words topic model with in-domain knowledge requirements.

G.1. Supervised Conditional Topic Random Field Model

Supervised Conditional Topic Random Field Model is created by J. Xu and E. Xing in 2010. as an attempt to utilize nontrivial input features in order to improve performance and to incorporate Markov dependency between topics assigned to neighbouring words .

This model presents further enhancement over Dirichlet Multinomial Regression and Conditional Topic Models in modeling using feature rich metadata, by employing a Markov dependency between topics thus accounting for sequential nature of textual data. This is accomplished through use of Conditional Random Field, a type of undirected graphical model.

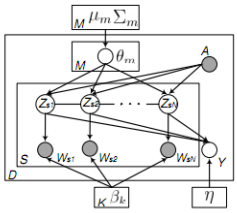

Figure 22. Plate notation of Conditional Random Field Model

H. Unsupervised sequence of words models with in-domain knowledge requirements

This class of topic models make use of word-order information, while attempting to increase applicability and interpretability of results to domain of interest with additionally supplied in-domain knowledge. This class is especially interesting because of lack of instances; there are few if any models that fall into this category. Authors are non-aware of such solutions.

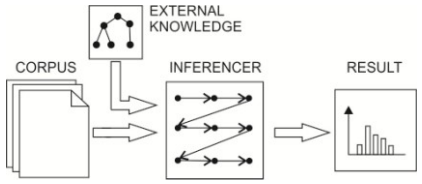

Figure 23. Outline of the unsupervised sequence of words topic model with in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Additional in-domain knowledge is supplied, usually in form of ontolgy or thesaurus. Word order is not neglected.

IV. CONCLUSION

A survey of most prominent probabilistic models is presented and a novel classification is proposed in order to emphasize fundamental approaches to probabilistic topic modeling. Motivation for such a survey rests in finding a new, possibly prospective direction of research. It is a fact that there are more models with simpler assumptions (bag of words models that do not use in-domain knowledge) than others. Bag of word models induce increased interest over sequence based ever since Latent Dirichlet Allocation was introduced. Models introducing in-domain knowlegde are sparse comparing to others, but with increasing need for semantically richer topics and applications, incorporating in-domain knowledge is gaining on popularity. Unsupervised models have generally precedence over supervised because of their applicability to large unlabeled corpora, but in problem domain where data labels exist, supervised models are presumably more reliable.

Perhaps the most notable contribution of this survey lies in observing the SDU class where none of the models are found. This observation may turn out to be the most valuable for doing further research.

REFERENCES

[1] A. Daud, J. Li, L. Zhou, and F. Muhammad, “Knowledge discovery through directed probabilistic topic models: a survey,” Frontiers of Computer Science in China, vol. 4, no. 2, pp. 280–301, Jun. 2010.

[2] David M. Blei. Introduction to Probabilistic Topic Models. Communications of the ACM, 2011 pp.

[3] Mark Steyvers, Tom Griffiths. Probabilistic Topic Models. In Landauer

[4] Zhu, Jun, and Eric P Xing. “Conditional Topic Random Fields.” *Forbes*. Ed. Johannes Fürnkranz & Thorsten Joachims.

[5] Steven Abney and Marc Light. 1999. “Hiding a semantic hierarchy in a markov model”. In Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing, pages 1–8.

[6] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, “Integrating topics and syntax,” In Advances in Neural Information Processing Systems 17, vol. 17, 2005, pp. 537–544.

[7] A. Gruber, M. Rosen-Zvi, and Y. Weiss, “Hidden topic markov models,” in Artificial Intelligence and Statistics, 2007.

[8] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. Suppl 1, pp. 5228–5235, Apr. 2004.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, Jan. 2003.

[10] X. Wang, A. McCallum, and X. Wei, “Topical n-grams: Phrase and topic discovery, with an application to information retrieval,” in ICDM 2007: Proceeding of the seventh IEEE International Conference On Data Mining, Ed., 2007, pp. 697–702.

[11] X. Wang and A. McCallum, “Topics over time: a non-Markov continuous-time model of topical trends,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD ’06. New York, NY, USA: ACM, 2006, pp. 424–433.

[12] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in Proceedings of the 23rd international conference on Machine learning, ser. ICML ’06. New York, NY, USA: ACM, 2006, pp. 113–120.

[13] D. Blei, T. Gri, M. Jordan, and J. Tenenbaum, “Hierarchical topic models and the nested chinese restaurant process,” 2003.

[14] D. M. Blei and J. D. Lafferty. (2006) Correlated topic models.

[15] T. Hofmann. (1999) Probabilistic latent semantic indexing.

[16] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” Journal of the American Society for Information Science, vol. 41, pp. 391–407, 1990.

[17] D. M. Blei and J. D. Mcauliffe, “Supervised topic models,”, in Proceedings of th Neural Information Processing Systems – NIPS, 2007.

[18] D. Mimno and A. McCallum, “Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression,” in Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI ’08), 2008.

[19] M. R. Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, “Learning author-topic models from text corpora,” ACM Trans. Inf. Syst., vol. 28, no. 1, pp. 1–38, Jan. 2010.

[20] D. M. Blei and P. J. Moreno, “Topic segmentation with an aspect hidden markov model,” in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR ’01. New York, NY, USA: ACM, 2001, pp. 343–348.

[21] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, “A Hidden Markov Model Approach to Text Segmentation and Event Tracking”, *Proceedings ICASSP-98, Seattle, May 1998*

[22] Y. Wang, H. Bai, M. Stanton, W. Y. Chen, and E. Y. Chang, “PLDA: Parallel latent dirichlet allocation for Large-Scale applications,” in Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management, ser. AAIM ’09, vol. 5564. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 301–314.

[23] W. Li and A. McCallum, “Pachinko allocation: DAG-structured mixture models of topic correlations,” in Proceedings of the 23rd international conference on Machine learning, ser. ICML ’06. New York, NY, USA: ACM, 2006, pp. 577–58