Please cite this paper as:

Khalifa, O, , Corne, D., Chantler, M., Halley, F (2013) Multi-Objective Topic Modelling, in : Evolutionary Multi-Criterion Optimization (EMO 2013), (Eds: Purshouse, Fleming Fonseca, Greco, Shaw), Springer LNCS, 15pp, 2013, to appear.

Multi-Objective Topic Modeling

Osama Khalifa, David Corne, Mike Chantler, Fraser Halley

Heriot-Watt University, Edinburgh
Corresponding author: dwcorne@gmail.com

Abstract

Topic Modeling (TM) is a rapidly-growing area at the interfaces of text mining, artificial intelligence and statistical modeling, that is being increasingly deployed to address the 'information overload' associated with extensive text repositories. The goal in TM is typically to infer a rich yet intuitive summary model of a large document collection, indicating a specific collection of topics that characterizes the collection - each topic being a probability distribution over words - along with the degrees to which each individual document is concerned with each topic. The model then supports segmentation, clustering, profiling, browsing, and many other tasks. Current approaches to TM, dominated by Latent Dirichlet Allocation (LDA), assume a topic-driven document generation process and find a model that maximizes the likelihood of the data with respect to this process. This is clearly sensitive to any mismatch between the 'true' generating process and statistical model, while it is also clear that the quality of a topic model is multi-faceted and complex. Individual topics should be intuitively meaningful, sensibly distinct, and free of noise. Here we investigate multi-objective approaches to topic modeling, which attempt to infer coherent topic models by navigating the trade-offs between objectives that are oriented towards coherence as well as converge of the corpus at hand. Comparisons with LDA show that adoption of MOEA approaches enables significantly more coherent topics than LDA, consequently enhancing the use and interpretability of these models in a range of applications, without any significant degradation in the models' generalization ability.

1 Introduction

Topic Modeling (TM) is a relatively recent and rapidly-growing field of research at the interfaces of text mining, artificial intelligence and statistical modeling; it is being increasingly deployed to address the 'information overload' associated with extensive text collections and repositories. The growing interest in topic modeling can be associated with the fact that text comprises about 85% of data worldwide [1]. Modern approaches to topic modeling are based on a variety of theoretical frameworks that tend to consider any individual document to be a weighted mixture of *topics*, where each individual topic is a multinomial distribution over words. An inferred *topic model* comprises a specific collection of topics, along with an assignment of one of these topics to each word in each document in the corpus at hand. Such a topic model can provide an efficient representation of the corpus and is effective at supporting a wide range of browsing and retrieval strategies (for example, delivering suitable documents in response to queries involving a weighted mixture of topics)[2].

Current topic modeling approaches such as Correlated Topic Models (CTM) and Latent Dirichlet Allocation (LDA), rely on finding a set of topics that maximizes the likelihood that the data were generated by a specific model of document generation. Though commonly returning interpretable results, the inferred models are ultimately aligned to a much-simplified abstraction of the real document generation process, and it is well understood that there is much room for improvement in terms of the intuitive coherence of the topics in relation to candidate 'real world' topics. In current approaches, it is therefore common to evaluate the inferred models via using them in a specific task such as classification of unseen documents. Such evaluation strategies are naturally partial, not representing a fully-rounded evaluation of a topic model. Meanwhile, such evaluation does not address the question of how more coherent topic models might be inferred in the first place.

A high quality topic model is one that can be expected to score well on a collection of different criteria, concerned with, for example, the coherence of individual topics, the coherence of the collection of topics as a whole, and the extent to which the inferred topics cover the entire collection, as well as the extent to which individual documents are explained by the topics (for example, a poor topic model in the latter respect may leave large portions of many documents unallocated to topics). However, each of these objectives is difficult to evaluate and can only be approximated – meanwhile, the familiar LDA likelihood criterion is a proven successful objective that, similarly, provides an appropriate and alternative approximate measure of quality. Exploiting the multi-criteria nature of topic models, in this article we begin to explore the use of multi-objective evolutionary algorithms (MOEAs) in topic modeling, and we investigate whether MOEA or MOEA/LDA hybrid approaches can be designed that yield better topic models than current approaches, and consequently provide enhanced effectiveness and user experiences in the many applications of TM technologies.

The remainder of the paper is organized as follows: in Section 2 we introduce concepts related to the most prominent topic modeling method, LDA, and we describe the topic model evaluation techniques that are generally used. Our MOEA approaches to TM are described in Section 3, and in Section 4 we describe a series of experiments that compare MOEA-TM approaches with LDA on three text corpora. Summary and final reflections are made in Section 5. Meanwhile at `http://is.gd/MOEATM` we provide source code, corpora and associated instructions that are sufficient to replicate our experiments and support further investigations.

2 Topic Modeling

Topic modeling is an approach to analyzing large amounts of unclassified text data [3]. It exploits the statistical regularities that occur in natural language documents in order to match queries to documents in a way that, though entirely statistical, carries strong semantic resonance. Good topic models should connect words with similar meanings (i.e. these words will typically co-occur within topics) and be able to distinguish between multiple meanings of a word depending on text context (i.e. the word 'set' will appear with high probability in a 'tennis' topic, and at the same time occur with high probability in a 'discrete mathematics' topic).

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is among the most prominent of current topic modeling techniques; it considers corpus documents to be underpinned by a mixture of latent topics, where each topic is characterized by a multinomial distribution over words [4]. LDA makes use of Dirichlet distribution which is a continuous multivariate distribution parameterized by a vector of positive reals. In the special case when all this vector's components are the same number, the distribution is called symmetric Dirichlet. A quick summary of LDA using LDA generative process is as follows. Let K be a pre-defined number of topics, $k \in [1..K]$ a number represents the topic, α a positive K-component vector, η

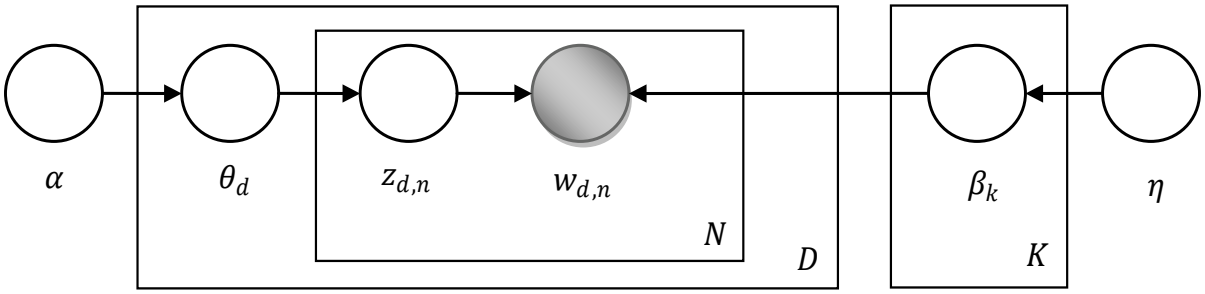

Figure 1: LDA model graphical representation of generative process.

a scalar, $Dir(\alpha)$ a K-dimensional Dirichlet distribution, V the corpus size, $Dir(\eta)$ a V-dimensional symmetric Dirichlet distribution, β_k a topic k distribution over corpus words, θ_d the topics proportion for one document, d a document from the corpus, w a word from the corpus, $w_{d,n}$ the n^{th} word in the document d, and $z_{d,n} \in [1..K]$ the topic assignment for the n^{th} word in document d.

for each *topic* k
  Choose a distribution over words $\beta_k \sim Dir(\eta)$
for each *document* d
  Draw a topic proportion $\theta_d \sim Dir(\alpha)$
  for each *word* w *in the document* d
    Draw a topic assignment $z_{d,n} \sim Multinomial(\theta_d), z_{d,n} \in 1..K$
    Draw a word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$

Figure. 1 shows a graphical representation of LDA model and shows the relation between latent and observed variables. The LDA generative process defines a joint probability distribution over the latent and observed variables as follows [5]:

$$P(\beta, \theta, z, w) = \prod_{k=1}^{K} P(\beta_k) \prod_{d=1}^{D} P(\theta_d) \left(\prod_{n=1}^{N} P(z_{d,n}|\theta_d) P(w_{d,n}|\beta, z_{d,n}) \right) \tag{1}$$

where, D is the number of documents, N the number of words inside one document, β is all topics distributions over corpus words, θ is the topics proportions for all documents, and z the topic assignments for all corpus words.

The main computational problem of LDA is to compute the posterior distribution – i.e., the conditional distribution of the latent variables given the observed variables. The posterior is given by the following formula:

$$P(\beta, \theta, z|w) = \frac{P(\beta, \theta, z, w)}{P(w)}. \tag{2}$$

Unfortunately, the exact posterior calculation is not feasible due to the denominator $P(w)$, calculation of which would involve summing the joint distribution over every possible combination of topic structures. However, there are various methods for approximating this posterior, such as Mean Field Variational Inference, Collapsed Variational Inference, Expectation Propagation and Gibbs Sampling [6]. LDA approaches that use Gibbs sampling are among the most popular methods in the current literature.

2.2 Evaluating Topic Models

The unsupervised nature of topic modeling methods makes choosing one topic model over another a difficult task. Topic model quality tends to be evaluated by performance in a specific application. However, other ways of evaluation are also used in the literature. Topic models can be evaluated based on *perplexity* [7] as a quantitative method; meanwhile, a 'human-evaluation' oriented evaluation method was introduced in [8] by creating a task where humans judge topics in terms of the frequency of apparently irrelevant words.

Perplexity is becoming a standard quality measure for topic models; it measures the topic model's ability to generalize to unseen documents after estimating the model using training documents. Lower perplexity means better generalization ability. Perplexity is calculated for a test corpus D_{test} by calculating the natural exponent of the mean log-likelihood of the corpus words [9] as follows:

$$Perplexity\,(D_{test}|\mathcal{M}) = e^{\frac{-\sum_{d\in D_{test}} \log P(w_d|\mathcal{M})}{\sum_{d\in D_{test}} N_d}} \tag{3}$$

where w_d represents words of test document d, $\mathcal{M}$ is the topic model, N_d is the number of words in document d. Thus, to evaluate two topic models estimated from the same training data, perplexity on test data is calculated, and the model with the lower perplexity value is preferred since it seems to provide a better characterization of the unseen data. However, perplexity does not reflect the topics' semantic coherence [10].

On the other hand, Pointwise Mutual Information (PMI) is an ideal measure of semantic coherence, based on word association in the context of information theory [11, 12]. PMI compares the probability of seeing two words together with the probability of observing the words independently. PMI for two words can be given using the following formula:

$$Pmi(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}. \tag{4}$$

The joint probability $P(w_i, w_j)$ can be measured by counting the number of observations of words w_i and w_j together in the corpus normalized by the corpus size. PMI-based evaluations correlate very well with human judgment of topic coherence or topic semantics [10, 13], especially when Wikipedia is used as a meta-documents to calculate the words co-occurrences within a suitably sized sliding window.

PMI values fall in the range $]-\infty, -\log P(w_i, w_j)]$, hence the higher the PMI value the more coherent the topic it represents. PMI values can be normalized to fall in the range $[-1, 1]$ as shown in [14] using the following formula:

$$nPmi(w_i, w_j) = \begin{cases} -1 & \text{if } P(w_i, w_j) = 0 \\ \frac{\log P(w_i)+\log P(w_j)}{\log P(w_i,w_j)} - 1 & \text{otherwise} \end{cases}. \tag{5}$$

The approach used to evaluate one topic is to calculate the mean of PMI for each possible word pair in the topic T. Consequently, the normalized PMI value for one topic T is given using the following formula:

$$nPmi_T = \frac{\sum_{w_i,w_j\in T} nPmi(w_i, w_j)}{\binom{T_{length}}{2}}. \tag{6}$$

where, T_{length} represents the number of words inside topic T.

3 MOEA Approaches to Topic Modeling

Optimization is the process of finding the best possible solution to a given problem under given limitations. For a single objective problem, the goal is to find the best solution that optimizes this objective e.g. the solution with lowest cost. However, most real-world problems are multi-objective, and multi-objective optimization aims to find a set of solutions that represent optimal trade-offs between the objectives. This is the set of *Pareto Optimal* solutions (after Vilfredo Pareto who introduced this notion [15]). There are a wide variety of approaches to multi-objective problems, however, many of these may fail when the Pareto front (the geometric structure of the Pareto set in objective space) is concave or disconnected [16]. Multi-objective Evolutionary Algorithms (MOEAs) tend to avoid these drawbacks [16, 17], among others, and are currently prominent among state of the art approaches to multi-objective optimization.

Topic models have many applications beyond unstructured text processing and text tagging. They can be used in analyzing genetic data [18], computer vision and object recognition [19], audio and music processing and speech recognition [20], emotion modeling and social affective text mining [21], and analyzing financial data [22]. Current topic modeling approaches such as LDA focus on producing topic models which score well on perplexity as measured over a test set of documents. However, other applications, such as text tagging which is used in digital libraries, requires highly coherent topics [10]. Considering the varied requirements of other applications of topic models, along with arguments made in Section 1, it is well-worth considering the investigation of MOEAs in an attempt to produce high quality topic models in general, and also in contexts relating to specific applications.

Our first approach to topic modeling using MOEAs (dubbed MOEA-TM) is to use an MOEA directly to optimize both perplexity and coherence (as measured by PMI). One element of the approach, in contrast to the prominent current methods such as LDA, is that we limit the number of words that can be allocated to a topic. This arguably has the benefit of leading to more readily interpretable topic models, but also has the significant advantage of reducing the computational load – in contrast, LDA (for example) does not limit the words per topic, and benefits from efficient techniques (such as Gibbs sampling) that are derived from the structure of the generating process. With this caveat, we take one of the state of the art MOEA approaches, so-called Multi-objective Evolutionary Algorithm Based on Decomposition (MOEA/D) [23], and adapt it to this task as described next.

3.1 Encoding and Generation of Initial Population

Each chromosome is a vector of topic variables $T_1, T_2.., T_K$ where K is the number of topics. Each topic variable contains a number of weighted words. Thus, each gene comprises two parts: the word index and a numerical value representing the word's participation in the topic. Chromosome structure is illustrated in Fig. 2. MOEA-TM can be used as a standalone approach to TM – i.e. starting from scratch to generate a topic model, or it can be used to enhance an existing topic model, such as one generated by LDA. In the standalone case, the population is initialized randomly as each topic variable is initialized on the basis of a randomly chosen document. Topic genes are initialized based on the most frequent words in the chosen document, with random weights. However, when the algorithm is used to enhance an existing model, the population echoes the model itself. Each topic variable is based on its corresponding model's topic, where the genes represent the highest weighted words in that topic. Eventually, in both cases Standalone and LDA-Initialized MOEA-TM the generated initial population contains number of solutions which all have the same number of topics and even same number of words inside each topic.

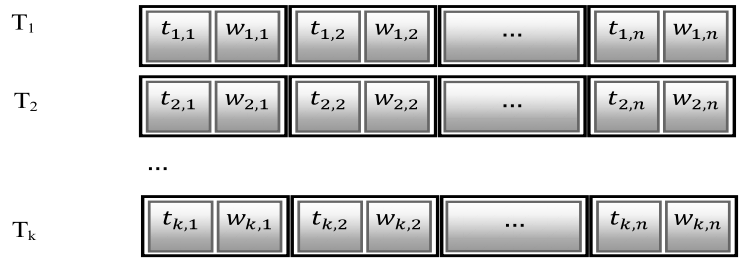

Figure 2: Chromosome Structure

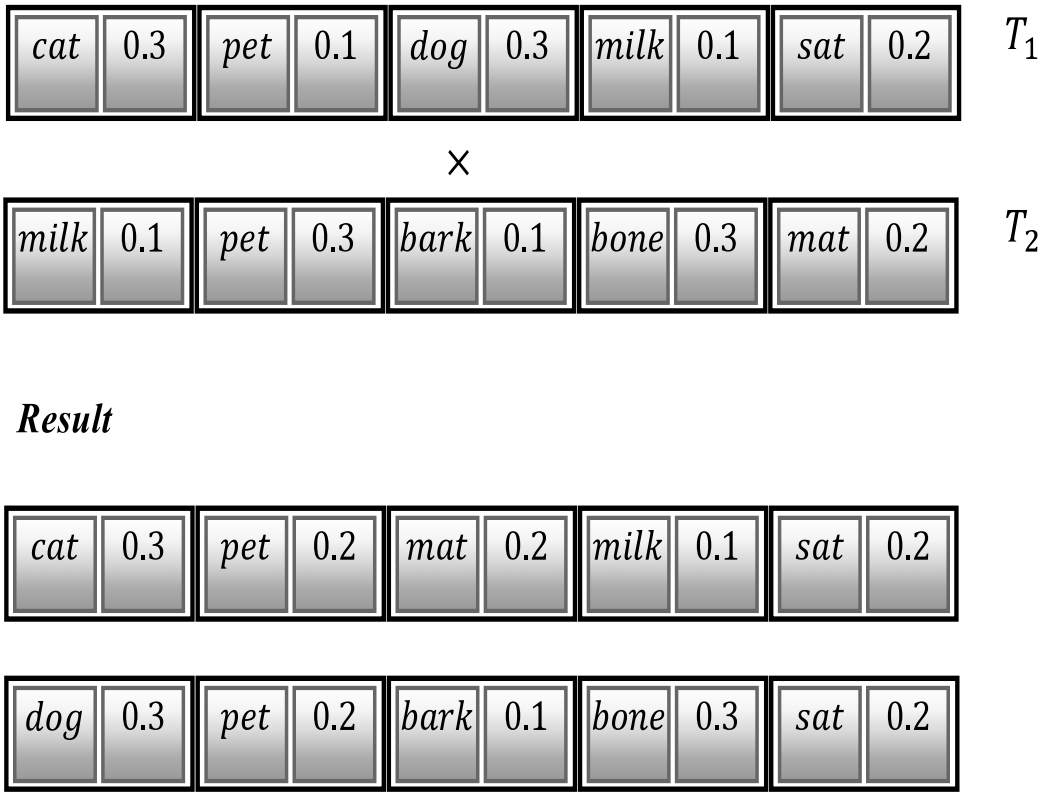

Figure 3: A simple two topics crossover example

3.2 Genetic Operators

Crossover in our current approach generates two offspring from two parents. Each child comprises as many topic variables as its parents has, via uniform crossover of the parents' corresponding topic variable genes, ensuring that words and their associated weights are copied together. However, when a word exists in both parents' topic variables, the children have the average word weight. A simple two topics crossover example is illustrated in Fig. 3.

Mutation is applied to a single randomly chosen gene, changing the weight to a new random number, and changing the word to another word from the corpus, ensuing that the newly introduced word occurs together in a document in the corpus with another randomly selected word from the topic variable.

3.3 Objectives

3.3.1 Coverage Score

This objective function applies pressure to encourage the topic models represent the whole of the corpus. For each document, the topics are evaluated by calculating the Euclidean distance between the weighted topics and the document itself. This is done by multiplying each topic's word-weight by the document's related topic weight, then calculating the distance between the resulting distribution and the original document word frequencies. Document-related topics weights are calculated using the following formula:

$$Prop_d(T) = \frac{\sum_{w \in T} tf_d(w)}{T_{length} - count_{w \in T,d}(w)} \tag{7}$$

where, $tf_d(w)$ gives the frequency of the word w in the document d and $count_{w \in T,d}(w)$ gives the number of words that exist in the topic and document at the same time.

Consequently, the coverage score for one document d can be given by the following formula:

$$Coverage_d = \sqrt{\sum_{w \in d} \left(tf_d(w) - \sum_{i=1}^{K} T_i(w) Prop_d(T_i) \right)^2} \tag{8}$$

where $T_i(w)$ gives the words weight if the word is existed in the topic T_i and zero otherwise. The coverage score can be normalized by dividing by its maximum value as follows:

$$nCoverage_d = \sqrt{\frac{\sum_{w \in d} \left(tf_d(w) - \sum_{i=1}^{K} T_i(w) Prop_d(T_i) \right)^2}{\sum_{w \in d} tf_d(w)^2}}. \tag{9}$$

This process is repeated for all corpus documents in order to calculate the whole topics coverage score over the corpus. Eventually, there will be a vector of values that need to be minimized. The overall score for corpus D is calculated by measuring the distance between the resulting vector and the center of the representing space using the distance formula:

$$CovObj = \sqrt{\sum_{d \in D} nCoverage_d^2}. \tag{10}$$

The objective $CovObj$ needs to be minimized in MOEA-TM algorithm.

3.3.2 Pointwise Mutual Information Score

This objective is responsible for optimizing the quality of the topics by increasing the topics' coherence using PMI. This is done by calculating the PMI score for each topic using (6). The PMI value need to be maximized because the higher PMI value the more coherent topics it represents. However, all objectives are minimized in MOEA-TM thus the original normalized PMI score for a topic T is substituted by the value $1 - nPmi_T$. This guarantees that minimizing PMI objective value represents high coherent topics. The overall score for whole topics is calculated by measuring the distance between the vector of topics PMI scores and the center of the representing space using the formula:

$$PmiObj = \sqrt{\sum_{i=1}^{K} (1 - nPmi_{T_i})^2}. \tag{11}$$

The objective $PmiObj$ needs to be minimized in MOEA-TM algorithm.

3.3.3 Perplexity Score

This objective is responsible for optimizing the model's ability to generalize to unseen data. Unfortunately, this score requires a topic model which assigns a topic to every word in the entire corpus. This means that this score cannot easily be calculated for a set of topics comprising only the most important corpus words, which is our approach in the 'standalone' MOEA-TM algorithm. Consequently, this objective is only investigated when MOEA-TM is used to enhance a pre-calculated topic model which is the case in our approach 'LDA-Initialized' MOEA-TM algorithm. Because perplexity value is natural exponent of a positive real – i.e. words negative log-likelihood mean, the negative log-likelihood mean is used instead. The Perplexity objective is calculated using the following formula:

$$PerpObj = \frac{-\sum_{d \in D_{test}} \log P(w_d|\mathcal{M})}{\sum_{d \in D_{test}} N_d} \tag{12}$$

where, $\mathcal{M}$ is the pre-calculated LDA topic model, D_{test} is a small test corpus, d document in the test corpus, w_d represents words of test document d, and N_d number of words inside document d. The $PerpObj$ objective is calculated using Left to Right method in [7] then normalized dynamically using other calculated values. The minimized negative log-likelihood mean leads to minimized perplexity; thus, this objective is minimized in MOEA-TM algorithm.

3.4 Best Solution

Because there is no topic modeling specific application applied in this paper, only one solution from the pareto front is compared with LDA. Extreme values are not considered although they provide better optimization for at least one objective. Thus, the solution in the middle is chosen and this is done by sorting the pareto front solution depending on a score representing the Euclidean distance between the objectives vector $\overrightarrow{v} = (v_1, v_2 \cdots v_n)$ and center of the objective space as follows:

$$score(\overrightarrow{v}) = \sqrt{\sum_{i=1}^{n} v_i^2}. \tag{13}$$

4 Experimental Evaluation

A number of experiments were performed to compare MOEA-TM with LDA, arguably the state-of-art in topic modeling. We used the LDA implementation with Gibbs Sampling which is provided by the MALLET package [24]. MOEA implementations utilized the MOEA Framework version 1.11 [25] run by JDK version 1.6 and CentOS release 5.8. Our evaluation uses three corpora: the first is a very small corpus with five documents created from Wikipedia and containing four rather distinct topics (Love, Music, Sport and Government). The second corpus is made from about 15000 documents taken from news articles covering mainly four topics: Music, Economy, Fuel and Brain Surgery. The Third corpus comprises about 800 documents that are summaries of projects in Information and Communication Technology (ICT) funded by the Engineering and Physical Sciences Research Council (EPSRC). Full details of each corpus are available from `http://is.gd/MOEATM`.

4.1 Standalone MOEA Topic Modeling

Standalone MOEA-TM was run ten times independently on each corpus, using only normalized coverage and normalized PMI objectives. LDA was also run ten times on each corpus. These experiments were done twice, once with number of topics set to 4, and once with number of topics set to 10.

Figure. 4, Figure. 5 and Figure. 6 show all MOEA-TM solutions resulted after 10 different runs. The average MOEA-TM Pareto Front is calculated by interpolating all MOEA-TM Pareto Fronts and

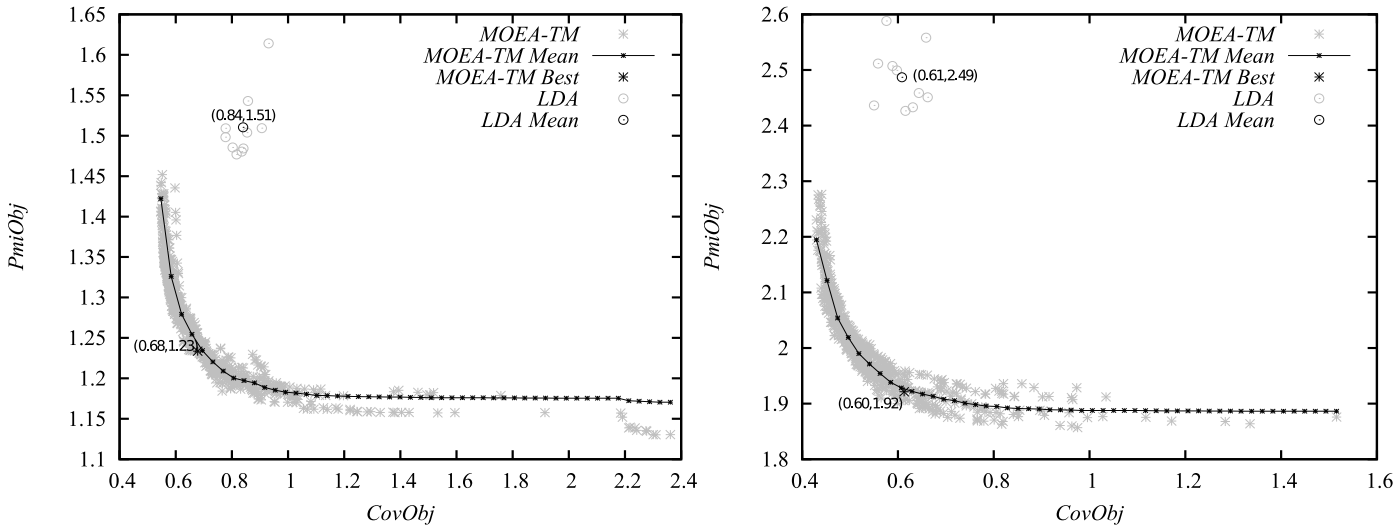

Figure 4: Wiki Corpus test: MOEA-TM Pareto Front and. LDA solutions for ten runs (average is taken), 4 topics left and 10 topics right. Lower is better

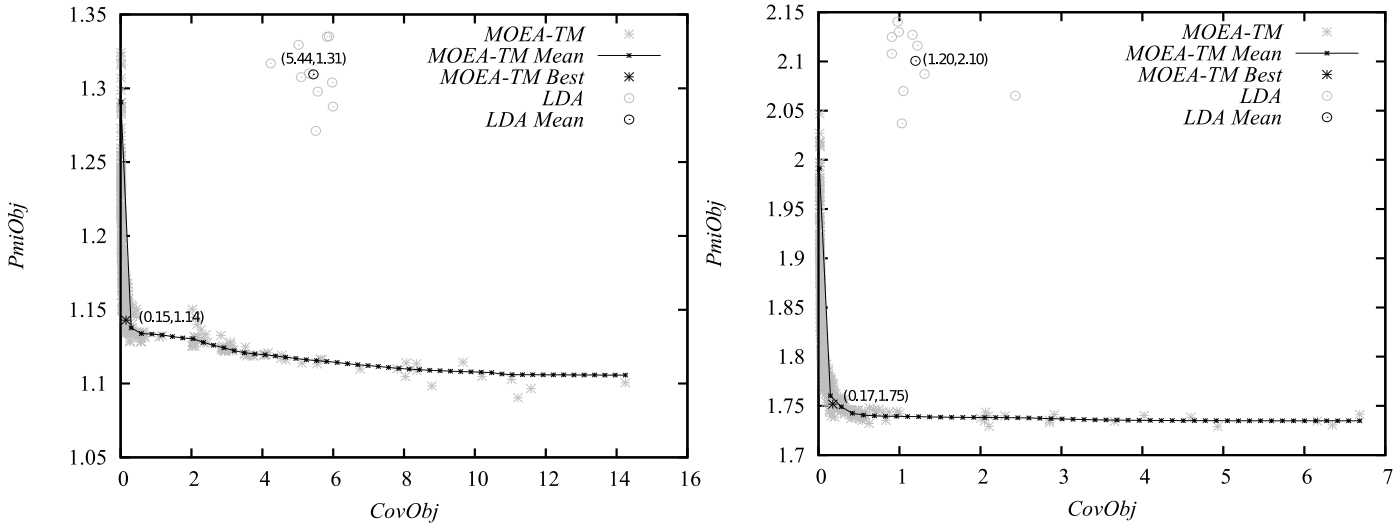

Figure 5: EPSRC Corpus test: MOEA-TM Pareto Front and LDA solutions for ten runs (Average is taken), 4 topics left and 10 topics right. Lower is better

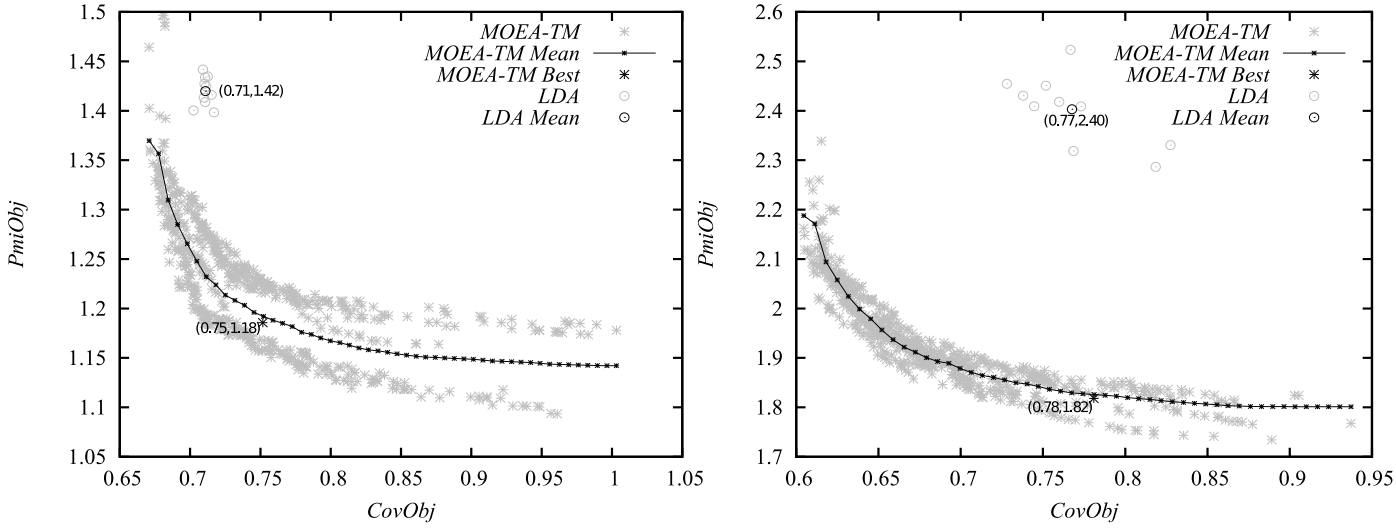

Figure 6: News Corpus test: MOEA-TM Pareto Front and LDA solutions for ten runs (Average is taken), 4 topics left and 10 topics right. Lower is better

then calculating the average Pareto Front curve. Best MOEA-TM solution, which is identified using (13), is displayed. LDA solutions and their mean values are shown in the figures as well. It can be seen that LDA is able to find relatively good solutions with an optimized coverage score; however the PMI (coherence) scores are poor in comparison to those found by MOEA-TM.

Figure. 4 and Figure. 5 show that best MOEA-TM solution optimizes both *PmiObj* and *CovObj* scores for the corpora Wiki and EPSRC respectively. On the other hand, Figure. 6 shows that for the News corpus MOEA-TM Best Solution was able to optimize the *PmiObj* but not the *CovObj* objective. This means that for this corpus LDA was able to find a higher representing topics but with poor PMI.

4.1.1 Evaluation:

Table. 1 and Table. 2 show the mean and sample standard deviations of original PMI metrics from the best MOEA-TM solutions and from LDA for 4 and 10 topic runs respectively. In these tables the higher PMI value is the better as the displayed values are the mean original normalized PMI values for solutions' topics after applying (6) over each topic.

Table 1: PMI scores for standalone MOEA-TM and LDA for the three corpora with four topics.

	MOEA TM		**LDA**	
	Mean PMI	**St. Deviation**	**Mean PMI**	**St. Deviation**
Wiki Corpus	0.3490	0.0128	0.2460	0.0194
EPSRC Corpus	0.4119	0.0091	0.3457	0.0102
News Corpus	0.3987	0.0178	0.2933	0.0082

It can be seen that MOEA-TM outperforms LDA in terms of the PMI metric. This means that topic models resulting from MOEA-TM are significantly more coherent than topics resulting from LDA. As suggested by the standard deviations, all MOEA-TM/LDA comparisons are significant with $p < 0.01$. The fact that MOEA-TM outperforms LDA in this respect is of course not very surprising given that LDA does not directly optimize PMI, however it is arguably surprising and interesting that the MOEA-TM approach can show such a marked improvement in topic coherence beyond that which seems achievable by LDA.

Table 2: PMI scores for standalone MOEA-TM and LDA for the three corpora with ten topics.

	MOEA TM		LDA	
	Mean PMI	**St. Deviation**	**Mean PMI**	**St. Deviation**
Wiki Corpus	0.3483	0.0078	0.2158	0.0163
EPSRC Corpus	0.4264	0.0080	0.3371	0.0106
News Corpus	0.3913	0.0077	0.2448	0.0216

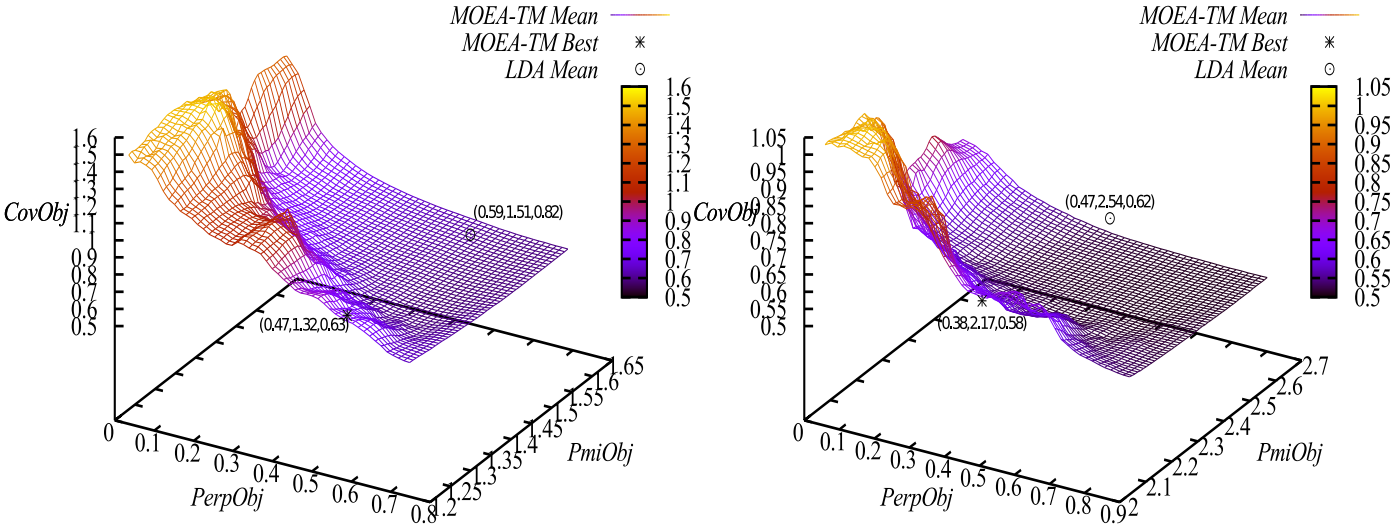

Figure 7: Wiki Corpus test: LDA-Initialized MOEA-TM Pareto Front and. Pure LDA solutions for ten runs (average is taken), 4 topics left and 10 topics right. Lower is better

4.2 LDA-Initialized MOEA Topic Modeling

In this experiment, similar experiments were run but in this case MOEA-TM is used to enhance a pre-calculated LDA topic model by optimizing three objectives *CovObj*, *PmiObj*, and *PerpObj*. The negative log-likelihood mean of an unseen test corpus words using the updated model is compared with the negative log-likelihood-mean of the same unseen test corpus words using the original LDA model. The model that has lower negative log-likelihood mean (or higher log-likelihood mean) is better as it leads to lower perplexity. LDA-initialized MOEA-TM was run ten times, and compared with (again) the results of ten unenhanced LDA topic models.

Figure. 7, Figure. 8 and Figure. 9 show the average MOEA-TM Pareto Front which is calculated by interpolating all MOEA-TM Pareto Fronts and then calculating the average surface. Best MOEA-TM solution, which is identified using (13), and LDA mean solutions are displayed in the figures. The MOEA-TM solutions and LDA solutions are not displayed for clarity. It can be seen that MOEA-TM was able to find better solution in terms of Coverage (*CovObj*) and PMI (*PmiObj*) for all corpora. In terms of perplexity (*PerpObj*) Figure. 8 shows that LDA was able to find better solutions for EPSRC corpus. Whereas MOEA-TM best solutions have better perplexity for Wiki and News corpora as shown in Figure. 7 and Figure. 9.

4.2.1 Evaluation:

Table. 3 and Table. 4 present the original normalized PMI and non-normalized negative Log-Likelihood metrics for LDA-Initialized MOEA-TM and LDA topic models with four and ten topics, respectively. It can be seen that LDA-Initialized MOEA-TM shows an improvement in terms of PMI values of 39%, 14% and 25% over pure LDA in the corpora Wiki, EPSRC and News, respectively when four topics are

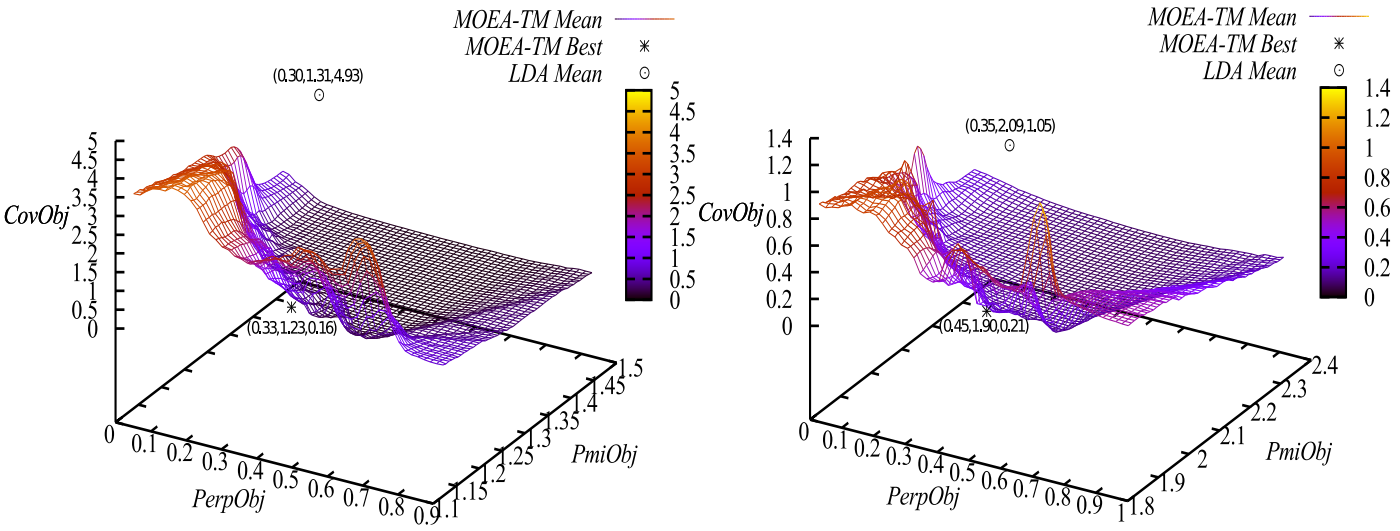

Figure 8: EPSRC Corpus test: LDA-Initialized MOEA-TM Pareto Front and Pure LDA solutions for ten runs (Average is taken), 4 topics left and 10 topics right. Lower is better

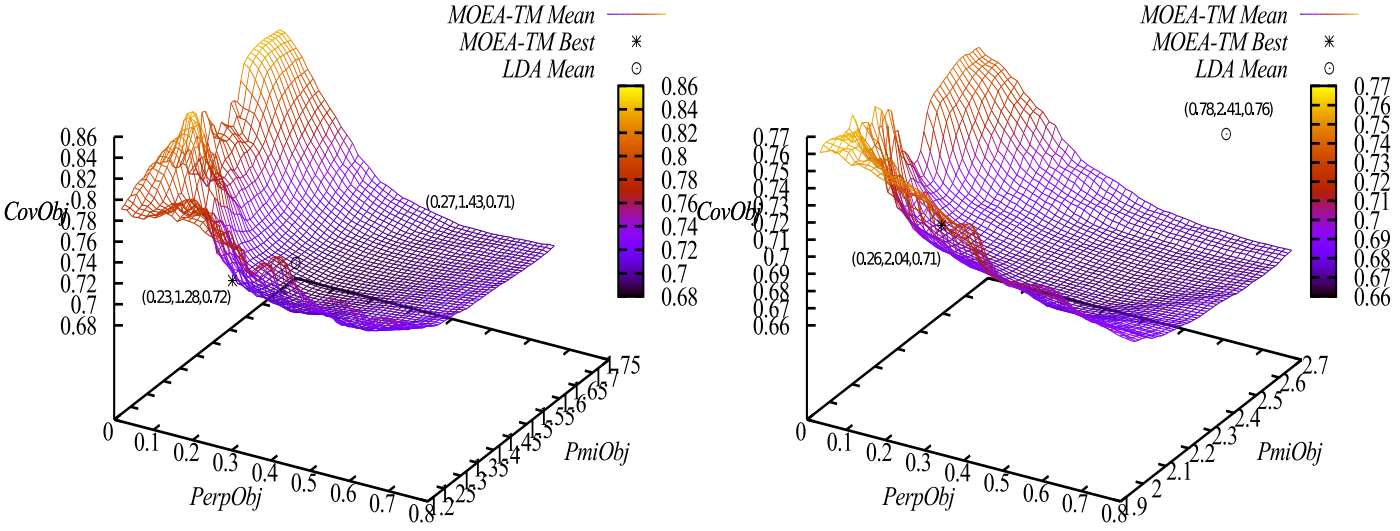

Figure 9: News Corpus test: LDA-Initialized MOEA-TM Pareto Front and Pure LDA solutions for ten runs (Average is taken), 4 topics left and 10 topics right. Lower is better

learned. When ten topics are learned the PMI improvement is 54%, 14% and 40% in the corpora Wiki, EPSRC and News, respectively. In all cases, a T-test again finds that the MOEA-TM improvement in PMI is significant with $p < 0.01$, while there is in contrast no significance in the difference in log-Likelihood values, suggestion that the improved coherence comes without any significant difference in the perplexity of the enhanced model.

Table 3: PMI scores for LDA-Initialized MOEA TM and Pure LDA for the three corpora with four topics.

	MOEA TM				**LDA**			
	PMI	**St. Dev**	**-LL**	**st. Dev**	**PMI**	**St. Dev**	**-LL**	**st. Dev**
Wiki Corpus	0.3443	0.1129	8.1417	0.0477	0.2476	0.1932	8.1488	0.0514
EPSRC Corpus	0.3933	0.0107	15.301	0.1128	0.3429	0.0094	15.293	0.1133
News Corpus	0.3653	0.0069	52.680	0.6756	0.2903	0.0142	52.835	0.7976

Table 4: PMI scores for LDA-Initialized MOEA TM and Pure LDA for the three corpora with ten topics.

	MOEA TM				**LDA**			
	PMI	**St. Dev**	**-LL**	**st. Dev**	**PMI**	**St. Dev**	**-LL**	**st. Dev**
Wiki Corpus	0.3105	0.0135	8.0716	0.0294	0.2013	0.0194	8.0822	0.0262
EPSRC Corpus	0.3889	0.0085	15.034	0.1005	0.3404	0.0101	15.096	0.0960
News Corpus	0.3428	0.0159	51.990	0.5377	0.2445	0.0208	53.261	0.6977

5 Conclusion

To sum up, MOEA-TM shows promising performance in topic modeling. MOEA-TM initialized from LDA models is able to enhance the coherence of the topic models significantly for each of the corpora tested here. A more coherent topic model is one in which the words that tend to appear together in a topic make more sense together to a human being. This can be very useful in many topic modeling applications, such as text tagging in digital libraries, where topic coherence is particularly important [10], while in general we would expect user confidence in inferred topic models, whatever the application, to be boosted when topics are coherent. In general, multi-objective approaches may contribute significantly to topic modeling, providing the ability to specify arbitrary objectives that may be relevant in a given application, and then providing the decision maker with a diverse collection of optimal models from which the most appropriate can be selected.

References

[1] Baars, H., Kemper, H.G.: Management Support with Structured and Unstructured DataAn Integrated Business Intelligence Framework. Inf. Sys. Manag 25(2), 132–148 (2008).

[2] Ha-Thuc, V., Srinivasan, P.: Topic Models and a Revisit of Text-related Applications. In Proceedings of the 2nd PhD workshop on Information and knowledge management (PIKM '08), pp. 25–32, New York (2008).

[3] Steyvers, M., Griffiths, T. L.: Rational Analysis as a Link Between Human Memory and Information Retrieval. In N. Chater and M Oaksford (Eds.) The Probabilistic Mind: Prospects for Bayesian Cognitive Science, pp. 327–347. Oxford University Press (2008).

[4] Blei, D.M., Ng A.Y. , Jordan M.I.: Latent Dirichlet Allocation. J. Mach. Learn. 3, 993–1022 (2003).

[5] Blei, D.M.: Probabilistic topic models. Commun. ACM. 55(4), 77–84 (2012).

[6] Srivastava, A., Sahami, M.: Text Mining: Classification, Clustering, and Applications (1st ed): Taylor and Francis Group, (2009).

[7] Wallach, H.M., Murray, I., Mimno, D.: Evaluation Methods for Topic Models. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1105–1112. ACM, Montreal Canada (2009).

[8] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading Tea Leaves: How Human Interpret Topic Models, in Advances in Neural Information Processing Systems, NIPS Foundation, Vancouver British Columbia (2009).

[9] Waal, A.d., Barnard, E.: Evaluating Topic Models with Stability, In: Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa, PRASA, Cape Town South Africa (2008).

[10] Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating Topic Models for Digital Libraries, In: Proceedings of the 10th annual joint conference on Digital libraries, pp. 215–224. ACM, Gold Coast Queensland Australia (2010).

[11] Su, Q., Xiang, K., Wang, H., Sun, B., Yu, S.: Using Pointwise Mutual Information to Identify Implicit Features in Customer Reviews, In: Proceedings of the 21st international conference on Computer Processing of Oriental Languages: beyond the orient: the research challenges ahead, pp. 22–30. Springer-Verlag, Singapore (2006).

[12] Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring Topic Coherence Over Many Models and Many Topics, In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952-961. Jeju Island Korea (2012).

[13] Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic Evaluation of Topic Coherence, In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108. Los Angeles California (2010).

[14] Bouma, G.: Normalized (Pointwise) Mutual Information in Collocation Extraction. In: Proceedings of The International Conference of the German Society for Computational Linguistics and Language Technology, pp. 31–40 (2009).

[15] Pareto, V.: Cours d'Economie politique. Revue Economique. 7(3), 426–430 (1896).

[16] Coello, C.A.C.: Evolutionary Multi-objective Optimization: a Historical View of the Field. Computational Intelligence Magazine IEEE. 1(1), 28–36 (2006).

[17] Coello, C.A.C.: Evolutionary Multi-objective Optimization: Basic Concepts and Some Applications in Pattern Recognition. In: Proceedings of the Third Mexican conference on Pattern recognition, pp. 22–33. Springer-Verlag, Cancun Mexico (2011).

[18] Chen, X., Hu, X., Shen, X., Rosen, G.: Probabilistic Topic Modeling for Genomic Data Interpretation. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 149–152. IEEE Press, Hong Kong (2010).

[19] Malisiewicz, T.J., Huang, J.C., Efros, A.A.: Detecting Objects via Multiple Segmentations and Latent Topic Models. Technical report, CMU Tech (2006).

[20] Smaragdis, P., Shashanka, M., Raj, B.: Topic Models for Audio Mixture Analysis. In: NIPS Workshop on Applications for Topic Models: Text and Beyond, Whistler Canada (2009).

[21] Shenghua, B., Shengliang, X., Li, Z., Rong, Y., Zhong, S., Dingyi, H., Yong, Y.: Joint Emotion-Topic Modeling for Social Affective Text Mining. In: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, pp. 699–704. IEEE Computer Society, Washington DC (2009).

[22] Gabriel, D., Charles, E.: Financial Topic Models. In: NIPS Workshop on Applications for Topic Models: Text and Beyond, Whistler Canada (2009).

[23] Zhang, Q., Li, H.: MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. IEEE Transactions on Evolutionary Computation. 11(6), 712–731 (2007).

[24] MALLET: A Machine Learning for Language Toolkit, `http://mallet.cs.umass.edu`.

[25] MOEA Framework: a Java library for multiobjective evolutionary algorithms, `http://www.moeaframework.rg`.