ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

Vol. 4, Issue 4, April 2016

# A Survey on Selection of Effective and Efficient Set of Reviews

Smita Patil[1], Dr. M.Z.Shaikh[2]

M.E. Student, Dept. of Computer Engineering, Bharati Vidyapeeth College of Engineering, University of Mumbai, Navi Mumbai, Maharashtra, India [1]

Principal, Bharati Vidyapeeth College of Engineering, University of Mumbai, Navi Mumbai, Maharashtra, India[2]

**ABSTRACT:** Online reviews are helpful for user to take appropriate decision while selecting a suitable product. However due the very long and irrelevant review content users face the problem to select the appropriate review. Now days Micro-Reviews are emerging as a new review concept in social media. Micro-Reviews are short (140-200 characters long) and highly focus on attributes and silent features of the product. Also these reviews can be posted by check-in services through mobile apps which make them more authentic and real time. This study proposes a novel mining problem which has used both the review sources (verbose review & Micro-Review) to select the small set of reviews which cover maximum attributes and silent feature of the product. This is two-step process: matching of review sentences with Micro-review and selection of small set of reviews that covers as many micro-reviews as possible. This problem is formulated by Integer Linear Programming and Greedy algorithm is used to obtain the optimal solution.

**KEYWORDS**: Micro-review; Efficient set of reviews; Effective set of reviews; Greedy algorithm

## I. Introduction

Now day's users find the ample of review data on various web sites about various products which gives the useful information to understand the product features and attribute. The selection of appropriate review from large corpora of the review is a big challenge which make more difficult by the verbosity and long length of the review. Reviewers often deviate by detailing personal stories that do not offer any useful information about the item or place being reviewed. This makes the task the harder to select the high quality and appropriate reviews.

Micro-Reviews and Micro blogging services are new emerging type of online reviews with recent growth of social media and networking sites. This new type of review content, which term as "Micro-reviews", can be found in micro-blogging services that allow users to "check-in", indicating their current location or activity. After checking in, a user may choose to leave a short message, up to 200 characters long, about their experience, effectively a micro-review of the place. Micro-reviews are alternate source of reviews which provide the useful information about product or place. Micro-reviews have several advantages:a) Micro reviews are short, distill and focused to most silent features of the product/place. b) Most of the Micro reviews written right when the users check in to the place, so they are spontaneous expressing reviewer's immediate and unadulterated reaction. c) Most authors check in by mobile apps, these authors are likely at the place when leaving the tips, which makes the tips more authentic.

Micro-reviews and reviews nicely complement each other. While reviews are lengthy and verbose, tips (Micro-Reviews) are short and concise, focusing on specific aspects of an item. At the same time, these aspects cannot be properly explored within 200 characters. This is accomplished in full-blown reviews which elaborate and contemplate on the complexity of a specific characteristic. In this study these two approaches are combined together to select the set of reviews which is small short and useful for the user. The concise and comprehensive reviews are more useful for the mobile applications of such sites, where the screen is limited, and the user attention span is shorter.In this study tips considered as a key data source to obtain the aspects of an item that the users care about, as well as the sentiment of the users. By covering the tips, effectively identify the review content that is important, and the aspects of the item upon which the reviews need to expand and elaborate. This is introduces as a novel formulation of review selection, where the goal is to maximize coverage while ensuringefficiency, leading to novel coverage problems. Efficiency constraint was not used in earlier review selection methods using maximum coverage. Also propose anInteger LinearProgramming (ILP) formulation, and provide an optimal algorithm. This allows quantifying the approximation quality of the greedy heuristics.In this study “Micro-Reviews” are referred as “Tips” and long verbose review as “Review”.

## II. Related work

In [1] there is complete search framework which, given a set of item attributes, is able to efficiently search through a large corpus and select a compact set of high-quality reviews that accurately captures the overall consensus of the reviewers on the specified attributes.it also introduce CREST (Confident Review Search Tool), a user-friendly implementation of our framework and a valuable tool for any person dealing with large review corpora. The efficacy of our framework is demonstrated through a rigorous experimental evaluation. In [2] Review summarization (generating statistical descriptions of review sets) sacrifices the immediacy and narrative structure of reviews. Likewise, review selection (identifying a subset of helpful or important reviews) leads to redundant or non-representative summaries. So to fill the gap between existing review-summarization and review selection methods by selecting a small subset of reviews that together preserve the statistical properties of the entire review corpus. It formalizes this task as a combinatorial optimization problem and show that it is not only NP-hard, but also NP-hard to approximate. It also designs practical and effective algorithms that prove to work well in practice. In [3] author describe about selection of comprehensive set of reviews, where Online user reviews are an invaluable resource for making informed decisions for a variety of tasks such as purchasing products, booking flights and hotels, selecting restaurants, or picking movies to watch. Sites like Yelp.com and Epinions.com have created a viable business as review portals, while part of the popularity and success ofAmazon.com and TripAdvisor.com is attributed to their extensive user reviews. The benefit of user reviews is that they are voluminous and comprehensive: multiple people, with different needs, different viewpoints, and different experiences review the same item, composing collectively a picture that is rich in detail and diverse in perspective. At the same time, this information abundance can be overwhelming to the users. In Amazon.com, for popular products such as digital cameras, there are typically several hundreds of reviews, many of which are fraudulent, uninformative, or repetitive. Typical online users do not have the patience to go through all of them to sort out the ones with useful information content. To address this problem, most online portals allow the users to rate reviews according to their helpfulness, and there has been substantial amount of research in automatically estimating the quality of a review. Such approaches produce a score for each review, or an ordered list of reviews. However, they do not account for the redundancy in the content of the reviews, or the fact that some important aspects of the reviewed item may not be covered at all by the top results. For example, all top reviews may be highly informative about the long range zoom of a new camera, but mention nothing about how easy it is to use, or to carry. Therefore consider the review set selection problem where given a set of reviews for a specific item, requirement to select a comprehensive subset of small size. The notion of comprehensiveness is defined with respect to the attributes of the product and the viewpoints of the reviews. Given a review of a specific item, and assumption is that we have the following information: (a) The attributes of the item that are discussed in the review; (b) The quality of the review; (c) The viewpoint of the review (e.g., positive or negative). The selected subset should cover as many attributes of the item as possible, while containing reviews of high quality, which offer different viewpoints for the attributes of the product. Then formalize this intuition as a maximum coverage problem, and show how we can extend existing algorithms for maximizing coverage to address our requirements.This method makes the following contributions: a. Formulate the review set selection problem as a coverage problem and define a generic formalism that can be used to model the different variations of our problem. b. It provides a theoretical analysis of the coverage problems and describes efficient algorithms for review selection. Whenever possible, provides approximation guarantees for the proposed algorithms.

## III. Problem statement and methodology

### A. *Problem Statement:*

Initial Ideal review set with perfect coverage and efficiency is rarely exist. In most cases most perfect efficiency is not essential. There may be some sentences which does not cover any of the tip but which make the review meaningful or enhance the readability of review. Hence review selection problem is formulated as Maximization problem where efficiency will be set as constraint and coverage will be maximize to obtain the optimal solution.

DOI: 10.15680/IJIRCCE.2016. 0404038

a. ($E_{FF}M_{AX}C_{OVERAGE}$) [4]-:*Select a set S of K reviews such that the coverage Cov (S) of the set is maximized, while the efficiency of the set is at least α, that is (Eff (S) ≥ α), and the size of the set S is minimized.*

In this formulation small set of review can be obtained by setting the desired number of K reviews.

b. ($E_{FF}S_{ET}C_{OVER}$)[4]-:*Select a set S R of reviews which covers all the tips in T, such that the efficiency of the set is at least α (Eff (S) ≥ α)and the size of the set S is minimized.*

This formulation gives the minimum number of reviews, which cover all the tips and understand the tradeoff between coverage and efficiency.

Where, R = a set of reviews, T = a set of tips, F = the matching function between review sentences and tips, α = efficiency of review and K = no of reviews.

B. *Methodology:*

Matching Reviews and Tips: Tips and reviews are of different granularity. Tips are short and concise where reviews are long and verbose. To match the review and tips, reviews is break in to sentences which are in semantic granularity similar to tips. The tips and review sentences will be match as per given above function and below three criteria.

a. Syntactic Similarity- In this criteria review sentences and tips which consist of similar words to be matched. Here sentences and tips will be considered as the bag of words. The sentences and tips which cover a similar word, it is considered that they convey the same meaning. Vector-Space model is well established model to find keyword similarity. Each review sentence “s” and tips “t” associated with vector “ṡ” and “ṫ” respectively. The dimensionality of the vector is the size of vocabulary. Each vector entry signifies the importance of the corresponding word. The degree of similarity between the sentence and tip is then measured as the cosine similarity.
SynSim (s, t) = cosine (ṡ, ṫ).

b. Semantic similarity- In this criterion the review sentences and tips which express the same meaning but using different words is matched. To match the semantic similarity between tips and review sentences, approach describe in this study is based on Latent Dirichlet Allocation (LDA) [5]. LDA associates probability distribution θt for each tip “t” over the topics, which captures the topics that are most important for t. Given the topics, and the corresponding language model for each topic as it is learnt from the tips, we can estimate the topic distribution θs for each review sentence s, which captures how well a sentence s reflects the topics being discussed in the corpus of tips. To measure the semantic similarity between a review sentence and a tip, we measure the similarity of the topic distributions θs and θt. A commonly used distance measure between two probability distributions is the Jensen-Shannon Divergence (JSD). Intuitively, a sentence and a tip are semantically similar if their topic distributions can describe each other well. The lower the divergence, the greater is the similarity.
Therefore, SemSim (s,t) = 1 – JSD(θs, θt)

c. Sentiment Similarity- in this criterion the review sentences and tips are matched based on opinion (positive or negative). The review sentences and tips which express the same opinion will be matched. A matching pair of sentences and tips should have same sentiment. To match the sentiment of the tips and review sentences Maximum Entropy Classifier (MEM) is used.

P(c+ | d) + P(c- | d) = 1
Where, d = given document (tips or sentences)
P(c+ | d) = conditional probabilities of MEM classifier outputs for positive classes
P(c- | d) = conditional probabilities of MEM classifier outputs for negative classes

The output of the MEM classifier, probability P(c+ | d) ϵ [0, 1], transformed in to polarity value polarity (d) = 2 P (c+ | d) – 1 ranging from -1 (extremely negative) to +1 (extremely positive). The polarity close to ½ is close to zero means document is neutral. Sentiment similarity between tips “t” and review sentences “s” are define as aproduct of their polarities. If the product of their polarity is 1 means they have similar polarity, if it is -1 then they have opposite polarity. If the product is zero means polarity is neutral.
SentSim (s.t) = polarity (s) x polarity (t)

d. Coverage Selection: The sentence “s” covers a tip “t” if they are matching pair of review sentence and tip. The review which covers at least one sentence which covers a tip we can say review “R” covers a tip “t”. The coverage can be defined for the set of reviews S which is sub set of Ṙ.

Cov (S) = |URϵS TR| / |T|

DOI: 10.15680/IJIRCCE.2016. 0404038

Where, Ṙ = collection of reviews R, S = subset of Ṙ, T = collection of tips, TR = set of tips which is covered by at least one sentence in review R.

a. Efficiency Selection: There are reviews which have high coverage (covering maximum number of tips) but they are too long containing many irrelevant sentences. To avoid the selection of this type of reviews concept of efficiency has been introduced in this study. Let consider Rt is the set of relevant sentences in the review which has total number of sentences “R”, so efficiency of the review can be written as Eff (R) = |Rt| / |R| Where, Eff (R) = efficiency of the review, Rt = number of relevant sentences in review which are matching with least one tip, R = total number of sentences in the review. By applying above efficiency to selection process each review with minimum threshold efficiency will be selected. This avoids the selection of verbose and very long review though they have high coverage but not matching the minimum threshold efficiency.

## IV. Algorithm

The review selection is from very large data corpus is a critical problem. To solve this problem greedy algorithm has been used which maximize the coverage to produce the optimal solution.The algorithm, shown in Algorithm 1, precedes in iterations each time adding one review to the collection “S”. At each iteration for each review R we compute two quantities. The first is the gain “gain (R)”, which is the increase in coverage that we obtain by adding this review to the existing collection S. The second quantity is the cost “cost (R)” of the review R, which is proportional to the inefficiency (1 - Eff(R)) of the review, that is, the fraction of sentences of R that are not matched to any tip. We select the review R_ that has the highest gain to cost ratio, and guarantees that the efficiency of the resulting collection is at least α, where α is a parameter provided in the input. The intuition is that reviews with high gain-to-cost ratio cover many additional tips, while introducing little irrelevant content, and thus they should be added to the collection.

Greedy EffMaxCoveralgorithm[4]:

Input: set of reviews Ṙ and tips T, efficiency function Fff, integer budget value K and parameters α, β

Output: A set of reviews S which is subset of Ṙ of size K.

```
S = θ
while |S| < K do
for all R ϵ Ṙ
        gain (R) = Cov (S U R) – Cov (S)
        gain (R) = β(1- Eff (R)) + (1-β)
end for
        Ɛ = {R ϵ Ṙ : Eff (S U R) ≥ α }
        if (Ɛ = = θ) or (maxRϵƐ gain (R) = = 0) then
break
end if
        R* = avgmaxRϵƐ gain (R) / cost (R)
        S = S U R*
        Ṙ = Ṙ \ R*
end while
return S
```

The cost of the review is parameterized by a value β ϵ [0,1], provided as part of the input, which controls the effect of efficiency in our selection of the review R_. More specifically, the cost of a review is defined as follows:

Cost (R) = β(1-Eff(R)) + (1-β)

When β = 0, the review selection is not affected by the efficiency of the reviews, but only by the coverage. For β close to 1 the effect of the efficiency on the review selection is maximized. Values in-between regulate the effect of efficiency in our selection. The higher the value of β, higher the value of coverage that is needed for a low efficiency review to be included in the set.

## V. Conclusion

This study has introduced the novel formulation of selection of reviews using micro reviews. The objective of increased coverage of micro reviews with limiting efficiency. The greedy algorithm is used to find the optimal solution with maximum coverage and minimum threshold efficiency.

## REFERENCES

1. T. Lappas and d. Gunopulos, “Efficient Confident Search in Large Review Corpora,” in proc. Eur. Conf. Mach. Learn. Knowl. Discovery databases: part ii, 2010, pp. 195–210.
2. W. Yu, R. Zhang, X. He, and C. Sha, “Selecting a diversified set of reviews,” in Proc. 15th Asia-Pacific Web Conf., 2013, pp. 721–733.
3. Thanh-Son Nguyen, Hady W. Lauw, , and Panayiotis Tsaparas, “Review Selection Using Micro-Reviews” IEEE transactions on knowledge and data engineering, vol. 27, no. 4, April 2015
4. E. Kouloumpis, T. Wilson, and J. Moore, “Twitter sentiment analysis:The good the bad and the omg,” in Proc. 5th Int. Conf. Weblogs Social Media, 2011, pp. 538–541.