J Intell Inf Syst
DOI 10.1007/s10844-014-0328-1

Algorithms and criteria for diversification of news article comments

Giorgos Giannopoulos · Marios Koniaris · Ingmar Weber · Alejandro Jaimes · Timos Sellis

Received: 10 June 2013 / Revised: 16 June 2014 / Accepted: 16 June 2014

Abstract In this paper, we introduce an approach for diversifying user comments on news articles. We claim that, although content diversity suffices for the keyword search setting, as proven by existing work on search result diversification, it is not enough when it comes to diversifying comments of news articles. Thus, in our proposed framework, we define comment-specific diversification criteria in order to extract the respective diversification dimensions in the form of feature vectors. These criteria involve content similarity, sentiment expressed within comments, named entities, quality of comments and combinations of them. Then, we apply diversification on comments, utilizing the extracted features vectors. The outcome of this process is a subset of the initial set that contains heterogeneous comments, representing different aspects of the news article, different sentiments expressed,

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

G. Giannopoulos (✉)
IMIS Institute, "Athena" Research Center, Athens, Greece
e-mail: giann@imis.athena-innovation.gr

M. Koniaris
School of ECE, National Technical University of Athens, Athens, Greece
e-mail: mkoniari@dblab.ece.ntua.gr

I. Weber
Qatar Computing Research Institute, Doha, Qatar
e-mail: iweber@qf.org.qa

A. Jaimes
Yahoo! Research, Barcelona, Spain
e-mail: ajaimes@yahoo-inc.com

T. Sellis
School of CSIT, RMIT University, Melbourne, Australia
e-mail: timos.sellis@rmit.edu.au

Published online: 16 July 2014

different writing quality, etc. We perform an experimental analysis showing that the diversity criteria we introduce result in distinctively diverse subsets of comments, as opposed to the baseline of diversifying comments only w.r.t. to their content. We also present a prototype system that implements our diversification framework on news articles comments.

1 Introduction

Over the last years the size of social web is growing exponentially. More and more users socialize through facebook, discuss current topics in forums, express their opinions/sentiments through blogs or twitter. The social web has also infiltrated in more traditional aspects of the web, such as news sites. Large corporations, like Yahoo! News,[1] allow their users to comment on news articles, facilitating the aggregation and public exposure of a wealth of user contributed information and opinions. Although this feature itself contributes largely to the spread of information and promotes the freedom of expression, data management issues come up due to the large amount of information to be handled.

It is often the case that news articles can gather tens of thousands of comments, which makes it impossible for interested users to review all of them. However, sometimes, the article's content itself is not enough for a user to form a complete view over a topic. The public opinion is a valuable resource that complements the article and represents the "wisdom of the crowds". In this case, the user needs to be able to review a very small amount of as heterogeneous as possible comments, that represent different aspects of the article. Upon that, the user is then able, by selecting a few initial comments, to further view "similar" comments to them, that is, focus on comments according to her specific preferences or personalized information needs. In this sense, heterogeneity (i.e. diversity) and personalization of information can be considered as two sides of the same coin. Both methodologies try to rerank result items (e.g. web pages or user comments), the former trying to capture all aspects of the information need and the latter trying to capture user preferences and restrict the results according to them. However, this does not mean that the use of one method excludes the other (Vallet and Castells 2012). For example, in an ideal search scenario, a user could be first presented with a small set of diverse results and, upon selection of some of them, the results could be personalized based on user's preferences. In a recommendation scenario, a state or the art recommender system that would mainly produce recommendations based on similarity of users or items, should occasionally introduce some heterogeneity on its recommendations, so that a user can identify new products or needs. And in our setting, a user might want first to review different aspects and opinions of a news article, so that she has a global, crowdsourced understanding of the topics discussed in it. Another use case scenario regards an archivist that needs to archive web information/resources about a specific topic. In this case too, the archivist should be able to "attach" to the primary resource (news article) complementary information (diverse comments). This process would help, e.g., a future journalist that tries to review past events, to gather as much diverse information on a topic as possible, in order to present an objective view on the topic.

In this paper, we propose a set of comment-specific diversification criteria to be applied for gathering heterogeneous sets of user comments on news articles. Diversification can be described as the selection of a subset of a result item set, which maximizes its heterogeneity w.r.t. specific criteria, and is a widely applied concept in several research areas (see

[1] http://news.yahoo.com/

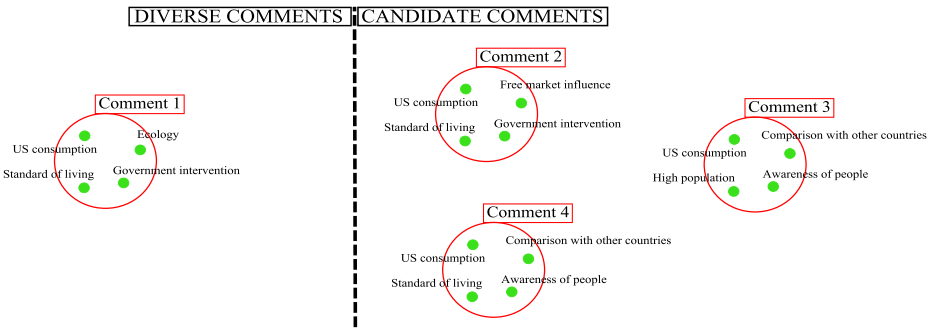

Fig. 1 Diversification of comments aims at diversifying the distinct information pieces (nuggets) within comments (Iteration 1) - the article regards US consumption rate

Section 3 for definition and analysis of diversification). Especially in the fields of web search and recommender systems, a plethora of works have been proposed (Drosou and Pitoura 2010), that handle several aspects of the problem. We claim that, although content diversity, namely plain textual diversity measured based on a "bag of words" model, suffices for other diversification settings (e.g. keyword search), it is not enough when it comes to diversifying comments of news articles. Thus, apart from content, with our proposed criteria we also capture sentiment expressed through comments, important Named Entities and writing quality of the comments. That it, we define criteria to capture semantic metadata of user comments, claiming that diversity of these semantic criteria induces diversity of topics/opinions/concepts described within comments. We implement the above criteria and we apply them on three state of the art heuristic diversification algorithms presented in Gollapudi and Sharma (2009), as well as on our proposed variation of diversification algorithm (*MAXSUM2*). To evaluate the effectiveness of our critiria, as opposed to plain content diversification, we extend the notion of *Information Nuggets*, defined in Clarke et al. (2008), so that it stands for news articles and comments. In short, we define as Information Nugget any possible topic/concept or sub-topic/concept found in the text of a news article or in its comments or any related interpretation/opinion/extension of the specific topics/concepts. So, the aim of the diversification process is to gather comments containing as many and as diverse Information Nuggets related to the topic of the article as possible. The following example illustrates this need.

In Fig. 1 four comments regarding a news article about "US consumption" are presented. We consider, for ease of presentation, that each comment contains four different information nuggets on the topic. Let comment 1 be already selected as the first comment in the diverse result set. The aim of the diversification process would be, then, to select the next candidate comment trying to maximize the heterogeneity of nuggets. In this case, comment 3 has 3 out of 4 different nuggets compared to comment 1, thus, being the most distant comment to it. Comment 2 has 1 out of 4 different nuggets, while comment 4 has 2 out of 4 different nuggets from comment 1. So, comment 3 is selected next (Fig. 2).

While in Fig. 1, where candidate comments are compared only to comment 1, comment 4 is more distant (diverse) to comment 1 than comment 2, things change in Fig. 2. Now that the result set contains both comment 1 and 3, comment 2 is the more distant, since it contains a nugget ("Freemarket influence") that is not present neither in comment 1 nor in comment 3. On the other hand, all nuggets of comment 4 are contained in the current result comments. So, comment 2 is the next one to be selected.

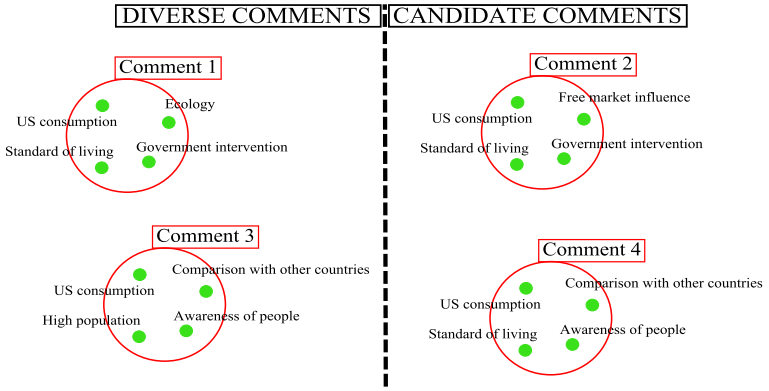

Fig. 2 Diversification of comments aims at diversifying the distinct information pieces (nuggets) within comments (Iteration 2) - the article regards US consumption rate

The above example illustrates how diversity of a comments set, w.r.t. a news article can be quantified and assessed through Information Nuggets extracted by the article and the comments. However, the task of recognizing concepts, topics and opinions is very difficult, even for a human evaluator. Even effective entity/topic/sentiment recognition tools, cannot recognize all concepts within a text passage or consistently recognize diverse representations of the same concept in several text passages. For example, when users include in their comments phrases such as: "Republicans", "conservatives" or "Bush's governance", depending on the article's context, they might refer to exactly the same concept. This is impossible for a machine to identify. The main contribution of this work are the user comments-specific diversification criteria we propose, that try to implicitly, however automatically, capture concept differences within comments. The diversification criteria we propose consider the following comment features. The first three were presented at our preliminary work on the problem (Giannopoulos et al. 2012), while the last three were added in the current work:

- Content (textual dissimilarity). This is the baseline diversity criterion that is also used in the rest of the literature to diversify search results. The objective is to obtain comments with diverse content.
- Sentiment. We consider the sentiment of users expressed in the respective comments, w.r.t. the news article content. Sentiment is measured in a nine grade scale ($[-4, 4]$), expressing negative, neutral, or positive sentiment. The objective is to obtain comments covering the whole range of sentiments.
- Named Entities (NEs). We consider the Named Entities (Persons, Organizations, Locations) found in the news article. Then, for each comment, we examine which of these NEs are referred in its content. Again, the objective is for the selected set of comments to contain as many article's NEs as possible.
- Sentiment on Named Entities. For each NE of a comment, we consider a window of words surrounding them and extract the sentiment only from the specific text area. This way, we focus the sentiment extraction only on the NEs per comment.
- Comment quality. Comment quality is measured in a seven grade scale w.r.t comment's readability. The objective is to select comments covering several readability levels.
- Aggregate comment quality. We consider, for each user, all the comments she posted on all the articles and produce an average commenting quality. Then each comment of the

user is represented by this quality score. The objective is to select comments covering the whole range of readability levels based on the aggregate user commenting quality.

We conduct a thorough experimental analysis that demonstrates the effectiveness of our methods, as opposed to diversifying comments only by content. The evaluation is performed on three measures we define in order to quantify the amount of Information Nuggets (total or distinct) found in different result comment sets, as well as the homogeneity of nuggets in comment sets. Almost all of the proposed variations consistently (and the best performing one significantly) outperform the baseline, achieving great differences in nugget coverage, in all evaluated diversification algorithms.

To sum up, our contributions are the following: (a) We define comment specific diversification criteria that go beyond plain textual (dis)similarity exploiting semantic metadata of comments, (b) We propose a heuristic diversification variation that performs very close to the best performing one and distinctively better than the other two state of art algorithms tested, (c) We extend the concept of information nuggets to the news articles/comments setting, defining intuitive evaluation measures to assess the effectiveness of the methods, (d) We perform a thorough evaluation of criteria and algorithms combinations, demonstrating the effectiveness of our methods and (e) We implement our methods into an initial prototype system that works on a publicly available news article dataset. To the best of our knowledge, the current, and our preliminary work in Giannopoulos et al. (2012), which is extended here, is the first work handling the specific problem. Also, our method is general enough so that it can be applied in other settings, such as comments on blog posts or forum discussions around a topic. The rest of the literature involves analysing user comments from several aspects, such as volume, political opinion, etc. (see Section 2). On the other hand, diversification is, in most works, handled from the aspect of diversifying search results.

The remaining paper is organized as follows. Section 2 presents related work and Section 3 discusses some background information on diversification objectives and algorithms. In Section 4, we present our method for diversifying news articles comments. Section 5 presents the implemented system. In Section 6, we present the experimental evaluation, that demonstrates the effectiveness of the proposed method and we discuss the experimental results. Finally, Section 7 concludes and discusses further work.

2 Related work

As stated in Section 1, to the best of our knowledge, there are no works that can directly be compared with our proposed method. In what follows, we present several approaches that deal with the problems of (a) news comments and, in general, social media analysis and (b) search results diversification.

The work in Wong et al. (2011) is the closest to ours. The authors present ongoing work on a system regarding online discussion groups. The system first requires that users explicitly state their opinions of specific topics. Then, it exploits this feedback to recommend several opinions, allowing the user to vary the similarity/diversity degree of the recommendations, w.r.t. her own opinions. Apart from the difference in the diversity criteria used, the system described in Wong et al. (2011) differs from ours in that it requires explicit, specific feedback from users and, also, it diversifies the recommended opinions w.r.t. each user's personal opinions and not in a global manner.

The authors of Li et al. (2010) propose a news recommendation system in forum-based social media, that exploits user comments to produce news recommendations. The

approach aims at building a topic profile, utilizing both the news text and its comments. This profile is then used to retrieve relevant news articles. Similarly, Shmueli et al. (2012) presents a method for recommending to users news articles that are likely to be commented by them. The authors propose a hybrid recommendation approach, where they exploit, apart from document content, the co-commenting patterns of users on the respective articles.

The authors of Tsagkias et al. (2009) first predict whether a news article is to receive any comments at all and, then, whether it will receive many comments or not. To this end, they apply two separate classification phases. In Tsagkias et al. (2010) they try to model and compare commenting distributions from several news sources and, also, predict comment volume by observing a short first period of commenting.

The work in Park et al. (2011) tries to capture commenters' sentiment patterns towards political news articles and to predict the political orientation from the sentiments expressed in the comments. The authors apply different learning techniques, depending on whether they predict political orientation for one or more commenters. They also take into account contextual information, such as the votes or links a comment received. In Munson and Resnick (2010) the authors study user comments on political news and evaluate readers' satisfaction on political opinions. In this way, they aim to differentiate between users who seek similar opinions to theirs and users who seek diverse ones. In Kucuktunc et al. (2012) sentiment analysis is performed in Yahoo! Answers user posts. The authors analyse the effect of several factors, such as demographics, topic, time on the expressed sentiments in users' answers.

The authors in Potthast (2009) study the descriptiveness of comments, i.e. the extent to which comments are similar to the topic they refer to. The authors obtain positive results, in the sense that a sufficient amount of comments can adequately represent the original commented text. In Diakopoulos and Naaman (2011) the authors perform a study on users' needs w.r.t. news article comments and conduct a quality analysis on comments posed in the articles of an online newspaper. In Herring et al. (2005) an analysis of links, comments and interconnections between blogs is performed. The authors of Hu et al. (2008) aim at producing document summaries, utilizing the respective comments. To produce the summaries, they extract sentences from the original document (e.g. blog post), which are biased to keywords extracted from the document's comments. In Mishne and Glance (2006) the authors perform an analysis on blog post comments and their relation to the posts. Specifically, they estimate the overall volume of comments in the blogosphere, analyze the relation between the weblog popularity and commenting patterns in it and measure the contribution of comment content to weblog access.

A thorough review of fundamental works in diversification is given in Drosou and Pitoura (2010). Carbonell and Goldstein (1998) describes the maximal marginal relevance method, which attempts to maximize relevance while minimizing similarity to higher ranked documents. To this end, the relevance of search results is calculated using two similarity functions, one measuring the similarity among documents, and the other the similarity between document and query. Chen and Karger (2006) consider an evaluation metric that penalizes a retrieval model only if it retrieves no relevant results at all. Given that, they propose a method where each result document is selected based on the probability that it is relevant to the previously selected ones.

In Gollapudi and Sharma (2009), the authors introduce a set of diversification axioms and show that it is not possible for a diversification algorithm to satisfy all of them. Also, they propose three diversification objectives. These objectives differ in the level at which the diversity is calculated, e.g. whether it is calculated per separate document or on the

average of the currently selected documents. The authors in Clarke et al. (2008) present a framework for evaluating novelty and diversity. Similarly, Agrawal et al. (2009) propose a greedy diversification algorithm but, also, extend some state of the art IR evaluation measures, so that they can be used in the context of diversification. Vee et al. (2008) presents a method for efficient diversification of structured data, where the items to be diversified are not documents, but objects with distinct attributes (i.e. records in a database table).

Finally, in Vallet and Castells (2012) the authors claim that diversification and personalization should not necessarility be considered antagonistic. They propose a series of methods that combine the two methodologies, by building personalization functionality on top of previously proposed, propabilistic diversification models. Their experimental evaluation, based on crowdsourcing, shows that combining personalization and diversification improves the precision of baseline ranking models.

As stated at the beginning of the section, none of the above works are directly comparable with are proposed framework. However, a common ground of most works on diversification, is that they base diversity on textual content of the items to be diversified. That is, their diversification objectives demand that result items have as diverse textual representation as possible. Thus, as baseline to compare our methods, we use a method that diversifies user comments only on their textual content (see Section 6).

3 Background

In this section, we give a short introduction of the concept of diversification and present a set of diversification objectives, algorithms and distance functions proposed in the literature.

3.1 The p-dispersion problem

The concept of diversification is closely related to the p-dispersion problem (Erkut 1990):

Definition 1 (p-dispersion problem) Select p out of n given points such that the minimum distance between pairs of the selected points is maximized.

The problem has several variations and different names (Chandra and Halldórsson 2001) such as *facility dispersion*, *p-defense*, *maxsummin dispersion*, etc. The problem is NP-complete (Erkut 1990), so a variety of heuristic algorithms have been proposed to solve the problem as efficiently as possible (Erkut et al. 1994). Next, we overview a categorization of these algorithms, based on the thorough study of Erkut et al. (1994).

3.2 Dispersion algorithms categorization

Dispersion algorithms can be divided into four categories: $construction$, $neighborhood$, $projection$ and $interchange$ algorithms. For what follows, let N be the set of candidate points, with $|N| = n$ and S the result set of diverse points, with $|S| = p$ when the diversification algorithm is completed.

Construction algorithms are divided into three subcategories:

– **Greedy construction heuristic** selects the two most distant points (according to a distance function) of the candidate set N to be inserted in the diverse results set S as an

initialization step. Then, until $|S| = p$, each time the next point to be inserted is the one that maximizes a distance w.r.t. to the points already inserted in S.

- **Greedy deletion heuristic** initially sets $S = n$ and then, at each step, removes one of the two closest points in S and specifically, the one that has the minimum distance to the rest of the points in S.
- **Semi-greedy deletion heuristic** is the same as *greedy deletion* with the only difference being that the selection of which of the two elements is eliminated is performed randomly.

Neighborhood algorithms consider (and add to S) an initial point x_i and define a neighborhood around it, i.e. a circle centered at x_i. Then they choose the next candidate point to be inserted in S, excluding all points in x_i's neighborhood. At each next step, the neighborhood is defined w.r.t. the lastly selected point x_i. These algorithms can differentiate w.r.t. the heuristic used to select the next point x_i' lying outside x_i's neighborhood: One can choose the *first*, the *closest* or the *furthest* point found to the points inside the neighborhood.

Interchange algorithms consider an initial random solution S and then, at each step, they interchange a point $x \in S$ with a point $y \notin S$, aiming to improve the objective function. Like *greedy deletion* the point x to be removed is one of the two closest points in S. The point y to be inserted may be the first one to improve the objective function or the one that improves the objective function the most. An important variation of this class of algorithms is *simulated annealing*, which aims at avoiding local maxima by periodically interchanging points that decrease the objective function value.

Projection algorithms project a Euclidean p-dispersion problem on a line and solve it there.

The evaluation performed in Erkut et al. (1994) shows that, in general, neighborhood and interchange algorithms are slightly more accurate than construction algorithms, while the projection algorithm has the worst performance. However, neighborhood algorithms need parameterization of the neighborhood radius and may not give a complete solution, in case of selection of large neighborhood radius. Also, they need to be run several times with shuffled point indices, so that the sequencing of candidate points does not bias the solution. Moreover, both neighborhood and interchange algorithms become time consuming for large N (number of candidate points). On the other hand, construction algorithms are easy to implement, parameter-free and perform relatively well, while being efficient for large sizes of N and small $|S|/|N|$ ratios. That is why a large number of recent works on search result diversification have adopted variations of greedy construction algorithms.

3.3 Objectives and distance functions

In this section, we present related work on search results diversification that has inspired us in part of our work, such as the definition of objectives and the evaluation methodology.

The authors of Clarke et al. (2008) base their analysis on the concept of information need u that is related to a query q and *information nuggets*. Information nuggets are different facets of the information need related to a query, such as different aspects or questions answered by the query's results. For example, for query 'jaguar', the obvious two individual nuggets would regard 'the car jaguar' and 'the animal jaguar'. Also, for query 'national elections', possible information nuggets would regard answering the following questions: 'when are the next national elections taking place?' or 'who are the candidate prime ministers in the next elections?' or 'what is the results estimation for the next elections?'.

Based on these definitions, they define the probability that a document d satisfies a query's information need u, as:

$$P(R = 1|u, d) = 1 - \prod_{i=1}^{m} (1 - P(n_i \in u) \cdot P(n_i \in d)) \quad (1)$$

where $P(R = 1|u, d)$ denotes that there is at least one information nugget n_i that satisfies the query's information need u and is covered by document d and m denotes the total number of information nuggets related to u. The product in the formula gives the probability that there is **no** nugget related to both u and d.

The above formula considers only one document, w.r.t. to its relevancy to the information need. So, it could be used as an objective to be maximized when we want to select the first item from a candidates set to insert into the diverse set S. At the general case, when there are k items inserted into S, (1) is transformed to:

$$P(R_k = 1|u, d_1, d_2, ..., d_k) = 1 - \prod_{i=1}^{m} (1 - P(n_i \in u) \cdot (\prod_{j=1}^{k-1} P(n_i \notin d_j)) \cdot P(n_i \in d)) \quad (2)$$

where $\prod_{j=1}^{k-1} P(n_i \notin d_j)$ denotes the probability that none of the previously inserted items (documents) satisfies nugget n_i. $P(R_k = 1|u, d_1, d_2, ..., d_k)$, eventually, gives the probability that the item k contains information nuggets that are not covered by the previously inserted items. Given the above, the objective of the heuristic greedy diversification algorithm is, at each step, to maximize the probability of (2), until S contains the desirable number of results.

The authors of Agrawal et al. (2009) extend the above analysis by considering categories to which queries and documents belong. Similarly to Clarke et al. (2008) they define $V(d|q, c)$ to be the quality value of document d, for query q, w.r.t. category c. Their objective is to find a set of k diverse results that maximizes the following quantity:

$$P(S|q) = \sum_{c} P(c|q)(1 - \prod_{d \in S} (1 - V(d|q, c))) \quad (3)$$

where $P(c|q)$ the probability distribution of categories for q.

The main difference of this method as opposed to Clarke et al. (2008) is that they take into account the relative importance between different nuggets, as well as that documents containing the same nugget may cover an information need to a different extend. For example, if a document d slightly refers to nugget n_i, then $V(d|q, c_i)$ is expected to be low, so other documents referring to nugget n_i might be necessary to be added at next steps, so that the information need is fully covered.

The algorithm adopted is a greedy construction algorithm where, at each step, the document with the highest marginal utility is inserted to S. The marginal utility is defined as:

$$g(d|q, c, S) = \sum_{c \in C(d)} U(c|q, S) V(d|q, c) \quad (4)$$

where $C(d)$ is the set of categories related to d and $U(c|q, S)$ is the conditional probability that the query q belongs to category c , given that all documents in S fail to satisfy the user information need. Essentially, the algorithm, at each step, seeks to find the document with the highest contribution in satisfying any of the query's categories, given the contribution of the previously inserted documents of S.

The authors of Gollapudi and Sharma (2009) incorporate into their objective the concept of document similarity. That is, their objective is to maximize a function that weights

the query-to-document similarity scores and the document-to-document distance scores. Three diversification objectives are considered: *Max-Sum*, *Max-Min* and *Mono-objective*. The objective functions are given below:

– **Max-Sum**

$$f(S) = (k-1)\sum_{u\epsilon S} w(u) + 2\lambda \sum_{u,v\epsilon S} d(u,v) \tag{5}$$

where S is the set diverse items, $|S| = k$ is the number of diverse items required, $w(u)$ is the similarity score of item u to the respective resource, $d(u, v)$ is the diversity score (distance) between items u and v and $\lambda > 0$ is a parameter specifying the trade-off between relevance and similarity.

– **Max-Min**

$$f(S) = \min_{u\epsilon S} w(u) + \lambda \min_{u,v\epsilon S} d(u,v) \tag{6}$$

– **Mono-objective**

$$f(S) = \sum_{u\epsilon S} w(u)' \tag{7}$$

where $w(u)' = (w(u) + \frac{\lambda}{|N|-1}\sum_{v\epsilon N} d(u,v))$ and N is the set of all candidate items.

Three approximation heuristic algorithms to maximize the above objectives are also proposed.

4 News comments diversification

In this section, we first define the problem we solve. Then, we present our proposed set of comments-specific diversification criteria, through which we try to capture the specific characteristics of comments on news articles and describe the implementation of four diversification algorithms we applied. Finally, based on the comments diversification criteria, we define the distance (for similarity and diversity) functions to be used by the algorithms.

4.1 Problem setting and definition

In this work, we consider the problem of returning k-diverse user comments for a news article. More specifically, the problem is formalized as follows:

Definition 2 (Comments diversification) Let A be a news article and N a set of comments on the article. Find a subset $S \subset N$ of comments that maximize an objective function f that quantifies the diversity of comments in S.

Most recent works on diversification consider the diversification of web search results w.r.t. a query. Our setting has several similarities to search result diversification, e.g. the fact that in both cases the items to be diversified are textual resources. Also, in both cases, along with diversity, one has to take into account the similarity of the resources to be diversified (results/comments) to the respective basic resources (query/news article). However, there are also substantial differences that impose the need of analysing and extending/adapting diversification algorithms and criteria, specifically on the setting of news article comments. Below, we briefly analyse the ones we consider more crucial:

- **Basic resource type.** Queries are short and, most of the times, they represent one or a few information needs which are, however, tightly related to the concepts corresponding to the keywords of the query. So, in the setting of keyword search, the diversification aims at distinguishing the different aspects-concepts of the query keywords and presenting a set of results that better cover these aspects. On the other hand, a news article contains much more text. Specifically, it consists of a complete and meaningful description of one or more topics, that may refer to partial subtopics. So, there is not a fixed, limited number of concepts to be diversified, as in keyword search. Also, the entities to be diversified might not even be simple concepts, but broader entities that consist of subconcepts (see motivating example of Figs. 1 and 2).
- **Diversification items type.** We safely assume that most top positions results returned by state of the art search engine models are at least somewhat relevant to the query posed. Most of them are expected to contain well structured text that clearly describes one or more topics related to the query, since the quality of these results has been assessed by the ranking functions of the respective search engines. On the other hand, user comments usually contain much less text and are of very diverse quality (missing punctuation marks, containing abbreviations or slang etc.). Also, some comments might be just replies to other user comments, or continuation of previous comments.
- **Opinion.** Most of the times, web results contain descriptions of concepts related to the query terms. On the other hand, most times comments express (to different extent) opinions and sentiments about the discussed concepts in the respective article.
- **Named Entities.** In the news article-comments setting, NEs are expected to play an important role. It is often the case that many of the concepts described within an article are related to one or more NEs and users are expected to refer to NEs, when commenting.

4.2 Diversification criteria

As stated in Section 1, most works on diversification measure diversity in terms of content, that is textual (dis)similarity between items. Even in works where more complex items are handled, e.g. Vee et al. (2008) where items to be diversified are records with attributes, again, the distinct diversification criteria are defined on the textual similarity or matching of distinct attribute values. In this section, we extend the notion of diversity on new dimensions (apart from content) that include sentiment, named entities and writing quality. Next, we describe the objective and the implementation of each criterion.

4.2.1 Content

We consider comments' content, which is the baseline diversification criterion, used in most works handling diversification, e.g. in web search results diversification. The importance of comments' content in the diversification process is straightforward. For each comment, we construct its term vector, with each feature corresponding to each distinct term found in the whole articles/comments corpus. Each feature value is computed by normalizing the term's frequency within the comment by the total number of terms the comment contains.

4.2.2 Sentiment

We consider the sentiment expressed by users through their respective comments. We propose that sentiment (positive, negative, neutral) is a diversification factor, since it expresses

users' opinions on the news articles' topics. In this sense, obtaining a set of comments that covers different classes of sentiment and, preferably, in a uniform manner, favors diversity.

We define nine classes of sentiment within the interval $[-4, 4]$, with -4 denoting very negative sentiment, 4 very positive sentiment and 0 neutral sentiment. We note here that we use nine classes due to the use of Sentistrength (Thelwall et al. (2010)) as a sentiment extraction tool. Thus, one could use arbitrary number of classes, without affecting the functionality of the method. However, the more sentiment classes are defined, the more refined results are expected, w.r.t. to the quality of the sentiment values that are extracted. Each comment is assigned two different characterizations w.r.t. to the sentiment expressed within it:

- **Maximum/minimum sentiment**. We consider the whole text of the comment. Out of this, we extract the maximum positive sentiment, as well as the minimum negative sentiment value.
- **Average sentiment**. We regard each sentence of the comment separately and extract the respective positive and negative sentiment values. Then, we take the mean average of these values for all the sentences of the comment.

The sentiment extraction process is based on specific words found in the comment's text that express positive/negative sentiment. The above distinction into two types of extracted sentiment is performed in order to capture different facets of the expressed sentiment/opinion. For example, a comment may contain only one sentence that includes a very positive sentiment regarding a specific subtopic (e.g. a specific person) mentioned in the news article. On the other hand, the rest of the comment might be, on the whole, negative towards all the other aspects of the article. With the distinction we propose, we are able to capture these differences in sentiment expression.

After the sentiment extraction process, for each comment, we construct two 9-feature sentiment vectors, one for each type of sentiment extraction, with each feature corresponding to a different sentiment class. Each feature takes a boolean value that denotes whether the specific sentiment class is expressed in the comment.

4.2.3 Named entities

We consider the Named Entities (NEs) found in the news article's text. These NEs might be Persons, Organizations or Locations. We suggest that NEs are important in terms of diversity, since news articles most of the times revolve around NEs. Even when an article talks about events or situations, usually one or more Persons or Locations are involved. Given that, it is important for a diversified comment set to cover as many article's NEs as possible.

For each of the aforementioned NE categories, we create distinct NE vectors, with each vector's features corresponding to the NEs found in the news article. For each comment, its feature values correspond to the frequency of the respective NE within the comment's text. In addition, we consider an aggregative NE vector that contains all NEs, irrespective of category. This results to 4 NEs vectors, that represent, for each comment, the coverage of article's Names Entities.

4.2.4 Sentiment on named entities

We consider the sentiment expressed around Named Entities in comments. That is, for each Named Entity, we consider a window of ± 5 words around it and extract the sentiment only

for the specific text area. The aim is to refine the sentiment extraction by focusing on the Named Entities, which are expected to be of greater importance than the rest of the terms in the comment's text.

We consider the NEs vectors defined in the previous criterion. Then, for each such vector, we define a new vector that assigns nine features for each NE, that is, one for each sentiment class regarding the specific NE.

4.2.5 Comment writing quality

We propose that comment writing quality is a diversification factor, since it expresses comprehensibility of the comment itself. Thus, it is important for a diversified comment set to cover as many different aspects of comment quality levels. Readability of a text represents the difficulty level of a written text through an numerical score obtained by applying a readability formula. Quantitative measures of text quality focus on characteristics of the text, such as word length and sentence length. The most influential of these is the Flesch Reading Ease Score[2] which combines number of syllables per word and average sentence length to produce a readability measure. The Flesch Reading Ease produces a score between 0 and 100 with higher values indicating easier texts. For each comment we apply a Flesch Reading Ease Score formula and assign a reading class to it. Then we construct a seven features vector with each feature corresponding to the seven distinct reading levels considered by the formula. Each feature takes a boolean value according to the grade level of the comment.

4.2.6 Aggregate comment writing quality

We regard each comment of the user separately and obtain the respective Flesch Reading Ease Score. Then we take the mean average of these values for all the comments a user has wrote and calculate a user comment writing quality level for each user. Finally, each comment of the user is assigned the average score and the respective feature vector.

4.2.7 User Co-commenting behavior

This criterion was finally excluded from our implementations and evaluation, due to poor effectiveness measured both in our previous work Giannopoulos et al. (2012) and in the current one. However, we briefly describe it, pointing out a possible cause of its ineffectiveness and consider it for further investigation in future work.

We consider the whole news article corpus. For each user, we construct a commenting vector, where each feature corresponds to a distinct news article. We suggest that the fact that a user comments on a news article (as well as the number of her comments on the same article) implies a relation between the user interests/opinions to the article's topics. Given that, the objective is to gather a diversified comment set, that corresponds to heterogeneous users. This heterogeneity, in our setting, is measured by the coverage of articles commented by the respective users. As mentioned above, this criterion, in its current implementation, marginally contributes to the diversification process, that is, it marginally affects the ranking of the comments. This is justified by the sparseness of the feature vectors for the specific criterion, since there are thousands of news articles, but most users usually comment on very few of them.

[2] http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php

4.3 Interpretation of objectives and implementation of diversification algorithms

Next, we describe the diversification heuristics we implemented in this work. That is, three state of art diversification algorithms presented in Gollapudi and Sharma (2009) (*MAXSUM1, MAXMIN, MONO-OBJECTIVE*) and the greedy Max-Sum variation we propose (*MAXSUM2*).

Max-Sum objective aims at maximizing the sum of all pairwise distances between items in S. Algorithms 1 and 2 give two approximations to solve the Max-Sum problem.

Algorithm 1 is presented in Gollapudi and Sharma (2009). The algorithm, at each step, examines the pairwise distances of the candidate items and selects the pair with the maximum pairwise distance, to insert into the set of diverse items S. Thus, the algorithm divides the whole set of items into diverse and candidates and works, at each step, **only on** candidates.

Algorithm 1 Produce diverse set of comments with MAXSUM1

Input: Set of candidate comments T, size of diverse set k
Output: Set of diverse comments S
$S = \emptyset$
for $i = 1 \rightarrow \lfloor \frac{k}{2} \rfloor$ **do**
Find $(u, v) = argmax_{x,y \in T} d(x, y)$
Set $S = S \cup \{u, v\}$
Set $T = T \setminus \{u, v\}$
end for
If k is odd, add an arbitrary document to s

In this work, we also implement Algorithm 2 for the Max-Sum problem approximation. We note that variations of the general logic of the algorithm might be implemented in several works (e.g. (Agrawal et al. 2009)). Differences lie in the initialization step and in the exact definition of the distance function. The algorithm initializes set S by inserting the most relevant comment to the articles. Then, at each step, it selects the candidate comment with the maximum distance to the centroid item of S.

Algorithm 2 Produce diverse set of comments with MAXSUM2

Input: Set of candidate comments T, size of diverse set k
Output: Set of diverse comments S
$S = \emptyset$
Find the most relevant comment u and set $S = \{u\}$
For any $x \in T \setminus S$, define $d_{MAX}(x, S) = d(x, c_s)$ where c_s the centroid of the comments contained in
S
while$|S| < k$ **do**
FIND $u = argmax_{x \in T} d_{MAX}(x, S)$
Set $S = S \cup \{u\}$
Set $T = T \setminus \{u\}$
Update c_s
end while

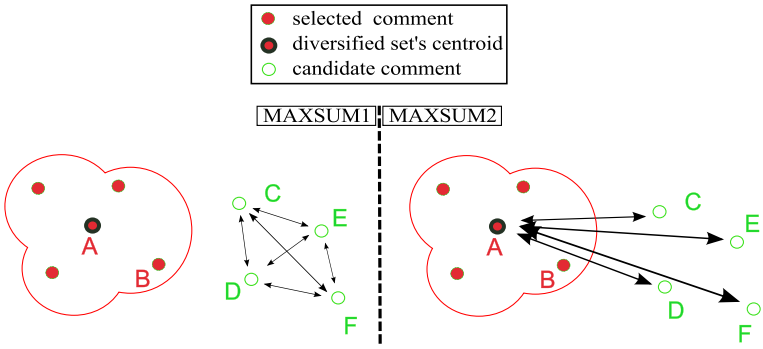

Fig. 3 Diversification algorithms 1 and 2

Although the algorithm aims at maximizing the same objective, its main difference with Algorithm 1 is that, at each step, it examines distances between candidate and already selected comments.

We graphically demonstrate the two algorithms' logic in Fig. 3. Consider the general case where a number of comments is already inserted in S (selected comments). The candidates are depicted as non-shaded circles in Fig. 3. The shaded circles represent already selected comments that constitute the *current* diverse comment set. *MAXSUM1* will examine all pairwise distances between candidate comments and result in selecting C and F, since they have the maximum pairwise distance. *MAXSUM2* will compare all candidate distances to A (centroid of S) and select the most distant one, that is F. Note that, in the next step, the centroid of S is re-calculated, so the distances of the candidate comments from the new centroid A' will change.

Max-Min aims at maximizing the minimum pairwise distance of the items in S. Finally, Mono-objective aims at simultaneously maximizing both similarity to the query and distance to the other documents, for each document belonging to S.

Algorithm 3 approximates the Max-Min objective. The algorithm (presented in Gollapudi and Sharma (2009)) initializes S with the same way as Algorithm 1. Then, at each step, it finds, for each candidate comment its closest comment belonging to S and calculates their pairwise distance d_{MIN}. The candidate comment that has the maximum distance d_{MIN} is inserted into S. Note that, in our implementation, we changed the initialization condition, so that it is consistent with Algorithm 2.

Algorithm 3 Produce diverse set of comments with MAXMIN

Input: Set of candidate comments T, size of diverse set k
Output: Set of diverse comments S
$S = \emptyset$
Find the most relevant comment u and set $S = \{u\}$
For any $x \in T \setminus S$, define $d_{MIN}(x, S) = \min_{u \in S} d(x, u)$
while $|S| < k$ **do**
FIND $u = argmax_{x \in T} d_{MIN}(x, S)$
Set $S = S \cup \{u\}$
Set $T = T \setminus \{u\}$
end while

Finally, algorithm 4 approximates the Mono-Objective. The algorithm, at initialization step, calculates a distance score for each candidate comment. The distance function weights each comment's similarity to the article with the average distance of the comment with the rest comments. After these distance scores are calculated, they are not updated after each iteration of the algorithm. So, each step consists in selecting the comment from the remaining candidates set with the maximum distance score and inserting it into S.

Algorithm 4 Produce diverse set of comments with MONO-OBJECTIVE

Input: Set of candidate comments T, size of diverse set k
Output: Set of diverse comments S
$S = \emptyset$
for each $x_i \in T$
Calculate $d(x_i) = w(x_i) + \frac{\lambda}{|T|-1} \sum_{v \in T} d(x_i, v)$
end for
while$|S| < k$ **do**
Find the candidate comment u = argmax$d(x_i)$
Set $S = S \cup \{u\}$
Set $T = T \setminus \{u\}$
end while

The logic of Algorithms 3 and 4 is depicted in Fig. 4. *MAXMIN* will examine, at each step, all pairwise distances between candidate comments and already selected comments. Then, it will select E as the comment with the maximum minimum distance to insert into S. Note that, in the next step, all distances have to be recalculated, since a new comment is inserted into S. *MONO-OBJECTIVE* will calculate, at initialization step, a score for each candidate comment and, then, it will start inserting comments into S, based on the initially calculated score. In our case, C will be the first one to be inserted, since it distinctively has the maximum average distance from every other comment. Note that, the scores are not updated.

The above examples demonstrate that, different diversification methods may produce different solutions, that is, different sets of diverse comments. Another important observation is that algorithms *MAXSUM2* and *MAXMIN* recalculate distances between candidate

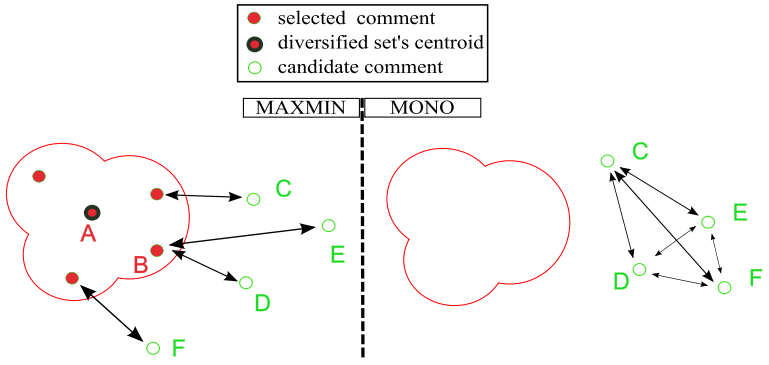

Fig. 4 Diversification algorithms 3 and 4

and already selected elements at each step. *MAXSUM1* also recalculates distances, but only between candidate items, and *MONO-OBJECTIVE* keeps the initially calculated distances through the whole process. Thus, we expect to observe differences in effectiveness when comparing these algorithms on the specific problem setting of diversifying news article comments.

Next, we describe the distance functions applied by each algorithm in our setting.

4.4 Scoring functions

Section 4.2 described the implementation of the proposed diversification criteria through the construction of criteria-specific feature vectors for each comment. The diversification algorithms utilize these feature vectors to calculate, at each step, an aggregate diversity score for each candidate comment. This score is, then, aggregated with the comment's similarity to the article to produce the final score for each candidate comment. The selection of the next result comment is based on this final score.

In order to produce a diversity score, we need to define a diversity function that measures the distance between two items. We adopt the widely used cosine similarity score and we define the diversity score of two items, u, v, w.r.t. a specific dimension i, as:

$$d_i(u, v) = 1 - \cos_i(u, v)$$

where $\cos(u, v)$ is normalized in the interval [0, 1]. We note that, the normalization is performed on the level of each criterion separately. That is, we calculate the maximum cosine similarity on each criterion and divide each similarity score with the maximum one, per criterion.

However, diversity is not the only objective: although the problem requires that heterogeneous comments are gathered, these comments ought to be relevant to the initial news article. So, the final score of each candidate comment, at each step, is a weighted sum of its relevance score to the news article and its diversity score. We define the relevance score of a comment u, w.r.t. the corresponding news article A, applying the cosine similarity measure on the article's and the comment's term vectors:

$$r(u, A) = \cos(u, A)$$

We note that this score is normalized in the interval [0, 1].

Depending on the diversification process we follow (as described in Section 4.3) we define four formulas that give the final score for each candidate comment u to be inserted into the set of diverse comments S, w.r.t. a news article A and an initial candidate comments set T:

$$score_{MAXSUM1}(u, v, A) = (1 - w) \cdot \frac{r(u, A) + r(v, A)}{2} + w \cdot \sum_{i=1}^{4} \lambda_i \cdot d_i(u, v)$$

where (u, v) is a pairs of comments, since this objective considers comment pairs for insertion, i is the diversification dimension, $w \epsilon [0, 1]$ is the weight of the total diversity score, as opposed to relevance score and $\lambda_i \epsilon [0, 1]$ is the weight of each individual diversity score, with $\sum_{i=1}^{4} \lambda_i = 1$.

$$score_{MAXSUM2}(u, A) = (1 - w) \cdot r(u, A) + w \cdot \sum_{i=1}^{4} \lambda_i \cdot d_i(u, C_i)$$

Table 1 Database schema

Table name	Description
article_data	Stores the text and metadata of the articles
comment_data	Stores the text and metadata of the comments
article_comments_dterms	Stores the distinct terms per article and its respective comments and the respective term frequencies for the article
comments_dterms	Stores term frequencies for the comments
article_cosine_vector	Stores the normalized term frequency vectors of the articles
comments_cosine_vector	Stores all diversification criteria in the form of feature vectors per comment
user_com_quality	Stores the user aggregated feature vector for the Aggregate Comment Writing Quality criterion

where C_i is the centroid of the current diverse set w.r.t. the diversification dimension i.

$$score_{MAXMIN}(u, A) = (1 - w)\cdot r(u, A) + w\cdot \sum_{i=1}^{4} \lambda_i \cdot d_i(u, \min v_{iu})$$

where $\min v_{iu}$ is the comment from the current diverse set that has the minimum distance to each candidate comment u.

$$score_{MONO}(u, A) = (1 - w)\cdot r(u, A) + w\cdot \sum_{i=1}^{4} \lambda_i \cdot \frac{1}{|T| - 1} \sum_{v \epsilon T} d(u, v)$$

5 System description

In this section, we provide some technical details of our diversification system and briefly describe the functionality of the system's GUI. We note that the described application is a desktop prototype that mainly helped us get intuition on the tested algorithms and criteria. One of our goals is the future implementation of a web based framework, for application on real web news portals, forums, etc.

We divide the diversification process in two stages: Offline and Online. Offline phase includes downloading raw news articles and comments data, preprocessing them to extract feature vectors for the diversification criteria, as well as term vectors for the relevance comparisons and storing them into the system's database. Online phase includes running the diversification algorithms on the extracted feature vectors. All data, before and after preprocessing, are stored in a relational (MySQL) database, the schema of which is presented in Table 1.

The system is implemented in Java. The data used in the specific process were downloaded from NY Times, using the respective APIs. For sentiment extraction, we use SentiStrength[3] (Thelwall et al. 2010) and for Named Entities recognition Stanford Named Entity Recognizer[4] (Finkel et al. 2005).

[3] http://sentistrength.wlv.ac.uk/

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

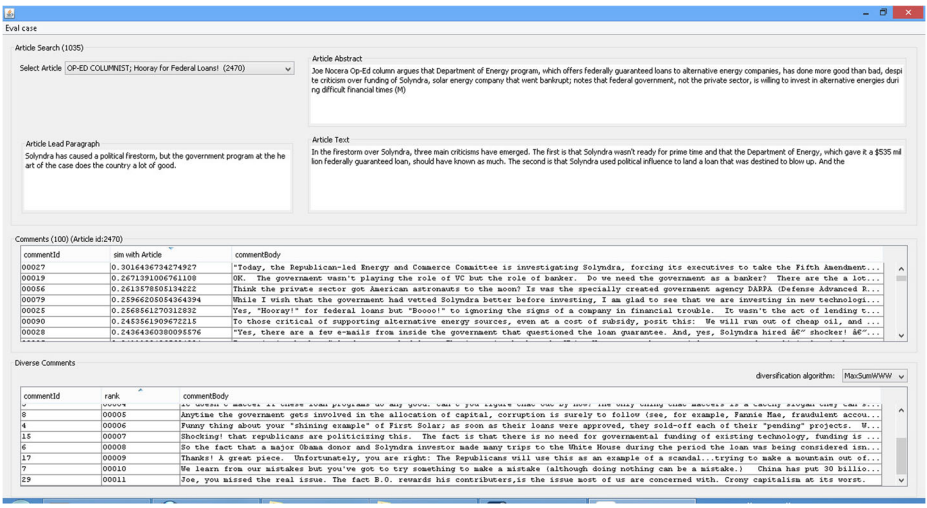

Fig. 5 System interface

Figure 5 presents a screen of the implemented prototype. Through the upper panel (”Article Search”) the user can select a news article and view the available information regarding it (text, abstract, lead paragraph). After an article is selected, its comments appear in the lower panels, sorted depending on the user choices. In the ”Comments” panel, all article’s comments are presented, sorted by date. In the ”Diverse Comments” panel, a set of diversified results are presented according to the user’s selection of algorithm and criteria combination.

6 Evaluation

6.1 Evaluated methods

In this section, we present a thorough evaluation of the proposed diversification algorithms and criteria. As a baseline to compare our proposed methods, we consider the naive (though state of art in other diversification settings such as keyword search) approach of *Content Diversity - CONTENTDIV*, which diversifies comments based only on their content. The rest methods are some distinctive variations of our approach, w.r.t. the combination of diversification criteria they utilize and are described next, along with the baseline:

- **Content Diversity - CONTENTDIV**. The baseline that applies diversification using only the criterion of Content diversity.
- **Sentiment Diversity - SENTIDIV**. The diversification variation that uses only the criterion of Sentiment.
- **Named Entities Diversity - NEDIV**. The diversification variation that uses only the criterion of Named Entities.
- **Sentiment around Named Entities - NESENTIDIV**. The diversification variation that uses only the criterion of Sentiment around Named Entities.
- **Hybrid Diversity - SEMIHYBRID**. The diversification variation that uses the criteria of Content, Sentiment and Named Entities, as presented in our previous work (Giannopoulos et al. 2012).

– **Extended Hybrid Diversity - HYBRID**. The diversification variation that uses the criteria of *Hybrid Diversity* combined with the extended criteria of Sentiment around Named Entities, Comment Writing Quality and Aggregate Comment Writing Quality, presented in the current work.

Each of the above variations was run for each of the four diversification algorithms presented in Section 4.3: *MAXSUM1, MAXSUM2, MAXMIN* and *MONO-OBJECTIVE*. We note that, when diversification criteria are combined, their scores are weighted equally to produce the final diversification score.

Also, we set a fixed weight for the diversity score to $w = 0.7$ and, thus, the weight for comment-to-article similarity to $(1 - w) = 0.3$. This way we wanted to ensure a minimum article-to-comment relevance guarantee, that would help to partially discard outlier comments. However, since the diversity score's weight is higher, and the scores are normalized, the relevance score does not significantly affect the results, as will be shown later in the evaluation.

6.2 Dataset

For our evaluation we produced a dataset of news articles and user comments from the New York Times. The online edition of the News Paper offers a well organized API to retrieve articles[5] and comments (The Community API[6]). Each resource, whether it is an article or a comment is accompanied by metadata, such as publication date, thematic categorization, user who posted it, etc., which are described in the respective API links.

In order to gather a sufficient amount of data we retrieved articles from the API, using the keyword "financial". This gave back around 2800 articles, for which we examined if there exist respective comments. Eventually, we obtained 1935 articles with a total of 293303 comments, which gives an average of 152 comments per article. We note that, since (a) the keyword we used is general enough and (b) it is searched on the whole text of the articles, the returned article set was not restricted to only financial articles, but it contained a wide range of topics, such as politics, business, economy, etc.

6.3 Evaluation methodology and metrics

The evaluation is performed as follows: We randonly select 10 articles to be evaluated, with the restriction that each selected article should have at least 100 comments. For each selected news article, we consider the total of its comments. For each of the six variations presented in Section 6.1 we return the top-10 result comments. Each of the six diversification variations, is applied in combination with each of the four diversification algorithms. So, in total, we evaluate 24 different methods.

The evaluation in our setting is based on the concept of information nuggets, which is presented in Clarke et al. (2008). In the query-results setting, nuggets may be different answers to question-queries, or different aspects of a topic/concept, for a query searching for information about a concept. In our article-comments setting, we adapt the concept of information nuggets:

[5] http://developer.nytimes.com/docs/read/article_search_api

[6] http://developer.nytimes.com/docs/community_api

Definition 3 (Information Nugget) Information Nugget is any possible topic/concept or sub-topic/concept found in the text of a news article or in its comments or any related interpretation/opinion/extension of the specific topics/concepts.

Of course the above definition is not strict enough and its application on specific news articles may give different results, depending on the strictness, the level of detail and the individual perception for topics of the annotator. However, it is a suitable definition for our experimental setting, since is allows us to map diversity into topics and concepts found in the textual descriptions of articles and comments, and, thus, to define proper diversity measures:

Measure 1 (Nugget Coverage - NC@n). With the first measure we quantify the extent of the nugget coverage from the result comments of each tested method. Basically, it is a Precision at N-based measure that measures how many nuggets are, in total, contained in the respective comment set. The measure is defined as follows:

$$NC@n = \frac{\sum_{k=1}^{n} I_k}{n \cdot |I|}$$

where n is the number of top comments, I_k the number of distinct information nuggets contained in comment k and $|I|$ the total number of distinct information nuggets. Since the maximum number of nuggets a comment may contain is $|I|$, thus, $|I_k| \leq |I|$, the measure's score is normalized within [0, 1].

Measure 2: (Distinct Nugget Coverage - DN@n). This measure is complementary to measure 1 and counts the ratio of the **distinct** nuggets found in the result set to the total of distinct information nuggets:

$$DN@n = \frac{\sum_{i=1}^{|I|} DFI_i}{|I|}$$

where DFI_i is defined as follows:

$$DFI_i = \begin{cases} 1 : FI_i > 0 \\ 0 : FI_i = 0 \end{cases}$$

where FI_i is the frequency of nugget i in the set of top-n comments.

Measure 3: (Nugget Uniformity - NU@n). With the third measure we try to quantify the heterogeneity of nuggets within comments, demanding that the nuggets are as uniformly distributed as possible. We define it as the variance of the nuggets' frequences in the result set from their mean value. If the mean value of nuggets frequences is defined as follows:

$$\bar{I} = \frac{\sum_{i=1}^{|I|} FI_i}{|I|}$$

then the nugget uniformity is defined as

$$NU@n = \frac{\sum_{i=1}^{|I|} (FI_i - \bar{I})^2}{|I|}$$

For each article, after all 24 methods are run, we consider the distinct set of the top-10 comments resulting from all methods. We keep only the comments' identifiers and we remove any provenance information, that is, which of the 24 methods produced each comment. Then, we separately apply the following two Information Nuggets extraction and annotation processes:

1) Manual Information Nuggets extraction
We execute the following two steps:

- **a) Information nuggets extraction**. After reading the news article and the respective distinct set of comments, we manually produce a set of Information Nuggets, w.r.t. the article's topics. This task is performed by two human Nugget Extractors. This way we create a pool of possible Information Nuggets for the article and its comments. We note that, for Nugget extraction we choose to also consider comments, apart from the news article, since, as stated in Section 1, we regard article's comments as valuable metadata that complement and enrich the article's information. We also note that this was not a linear process: the Extractors were instructed to go back and re-examine the article and the already reviewed comments, in case they discovered the need for e.g. definition of a new Nugget, the need to merge two Nuggets, etc. Thus, each article and its set of total distinct comments have been run through several times, in order to gather a complete and representative set of information nuggets.
- **b) Comments annotation**. Then, we assign two external Judges/Annotators the Nugget annotation process: We present them (i) the article's text, (ii) the distinct comments set produced from the results of all methods, again, stripped off of provenance information to avoid influencing a Judge by, e.g. observing a consistent pattern for a specific method, and (ii) the set of Information Nuggets. The Judges are asked to annotate each comment with the Nuggets they believe are represented through the comment. Once we obtain annotations for all comments, we apply the previously described evaluation measures. Note that the Extractors are different persons than the Judges, since we do not want to burden annotators with the task of defining Information Nuggets. Annotators are instructed to select Nuggets only included in the pool created by step (a). This separation between Extractor and Judges also ensures the objectivity of the process.

  Essentially, Information Nuggets is an auxiliary concept we define in order to quantify diversity. The aim is to disintegrate the generic concept of diversification into concise, minimum units of information, each one representing a different aspect of the topics of the article. This, way, diversification can be quantified and better assessed by the Annotator/user; instead of asking the Annotator to grade in general the diversity/novelty/interestingness of a set of comments, we ask her just to annotate comments with a pre-extracted set of Information Nuggets. This process (a) reduces the complexity of the annotation task, (b) captures the purpose of the task utilizing Information Nuggets which express diversity and (c) makes the evaluation task independent of the criteria used: while for diversity critiria we use, among others, the raw text of the comments and identified Named Entities, as Information Nuggets we use user identified concepts, that might not correspond to actual words or phrases from the comments' text. So, essentially, we utilize Information Nuggets as user-annotator feedback to measure the effectiveness of the methods.

2) Automatic Information Nuggets extraction

We also tried an automatic approach for extracting information nuggets from the article and its comments. Specifically, we used OpenCalais Web Service,[7] a tool which automatically creates rich semantic metadata for user submitted content. OpenCalais analyzes the submitted text and extracts Named Entities, Facts and Events within it. We also used

[7] www.opencalais.com

Table 2 Evaluation articles and indicative Information Nuggets (manually extracted)

Article's main topic	Indicative nuggets
Tax evasion	Tax evasion, Ethics, Law and Legislation, Politics
Obama's anti-foreclosure plan	Housing bubble, Politics, People's irresponsibility, Mortgage crisis
Scandal with politician and bankers	Bailout, Bank's name, Legislation, Corruption, Discrimination
Financial reform related to elections	Goldman Sachs, Subprime mortgage crisis, Economic ideologies, Wall Street
Federal loans on energy programs	Alternative energy, Politics, Competitiveness, Financial crisis, Political parties
US consumption rate	Consumption comparisons, Free-market economy, Solutions, Self criticism
Democrats nominations	Clintons unite Republicans, Obama represents change, Critisism on candidates
Prescriptions decrease: consequences/reasons	Critisism on corporations, Economical drug solutions, Patients examples
Obama measures on financial crisis	Critisize irresponsible americans, Blame free market, Measures are moderate
Relation between successful people/elite colleges	Community college inferior to others, How students exploit education matters

AlchemyAPI[8] which also analyses text and extracts abstract concepts. Although the resulting Information Nuggets from the two tools are (expectedly) poor compared to the more meaningful Nuggets extracted by human Extractors, we also performed expectiments on them to obtain more evidence on the generality of the evaluation results.

In Table 2 we present the general topics of the 10 evaluated articles and some general (randomly selected from both the article's text and its comments), indicative Information Nuggets that were manually extracted. In Table 3, respectively, we present some indicative automatically extracted Nuggets from the the articles and their comments. The information presented in these two Tables indicates what we carefully deduced by examining/comparing the sets of Nuggets for each article: although there are similarities on some of the Nuggets between the manual and the automatic extraction setting, automatically extracted Nuggets are strictly based on terms and phrases identified within the text of the comments. Thus, some of them might be irrelevant of the topics of the article or might overspecialize a concept in such a degree that makes it impossible to identify the concept in other comments (although they implicitly contain it). Thus, we consider the evaluation results on the manual Nuggets extraction setting (Section 6.4.1) more reliable than the ones on the automatic setting (Section 6.4.1).

[8] http://www.alchemyapi.com/api/concept/

Table 3 Evaluation articles and indicative Information Nuggets (automatically extracted)

Article's main topic	Indicative nuggets
Tax evasion	Taxation in the United States, Barack Obama, Moonlight, Banking in Switzerland
Obama's anti-foreclosure plan	United States housing bubble, Payments, Labor, Hudson River, Price, Receipt
Scandal with politician and bankers	Law_Crime, California, Subprime mortgage crisis, Military personnel, Question
Financial reform related to elections	Subprime mortgage crisis, Illinois, Long-Term Capital Management, Good, Vaccination
Federal loans on energy programs	Solar panel, Hydrocarbon, Personal finance, Energy industry, Chevrolet
US consumption rate	Political repression, Environmental issues, Broadsheet, Wheat, Question
Democrats nominations	North Carolina, Joe Biden, Southern hip hop, Criticism, President of the United States
Prescriptions decrease: consequences/reasons	Drug rehabilitation, Americas, Stroke, Public choice theory, Moon
Obama measures on financial crisis	Economics terminology, Taxation in the United States, Newspaper, Real estate bubble
Relation between successful people/elite colleges	Student financial aid, SAT, Emotion, Committee on Institutional Cooperation

A more detailed description of the articles and the corresponding Information Nuggets is given in the Appendix. Specifically, in the Appendix tables, for each of the ten evaluation articles, we provide its main topic, its abstract and the Information Nuggets manually extracted *only* on the article's abstract. Due to excessive space requirements, we omit the respective comments and the information nuggets corresponding to them. However, we believe that the provided tables in the Appendix suffice to provide an intuitive view of the use of Information Nuggets in our experimental evaluation.

6.4 Evaluation results

Next, we present the two evaluation settings we followed. The first, and more important in terms of reliability, is based on manually extracted Information Nuggets (from human Extractors) that where used by separate human Annotators to annotate comments. The second setting is based on automatically extracted Information Nuggets from OpenCalais and AlchemyAPI services. We note that we report on combined results from the two aforementioned automatic topic extarction tools.

6.4.1 Measures on manually extracted information nuggets

In Fig.s 6 and 7 we present the Nugget Coverage and Distinct Nugget Coverage values respectively, for comment result positions from 1 to 10. We note that the two measures are

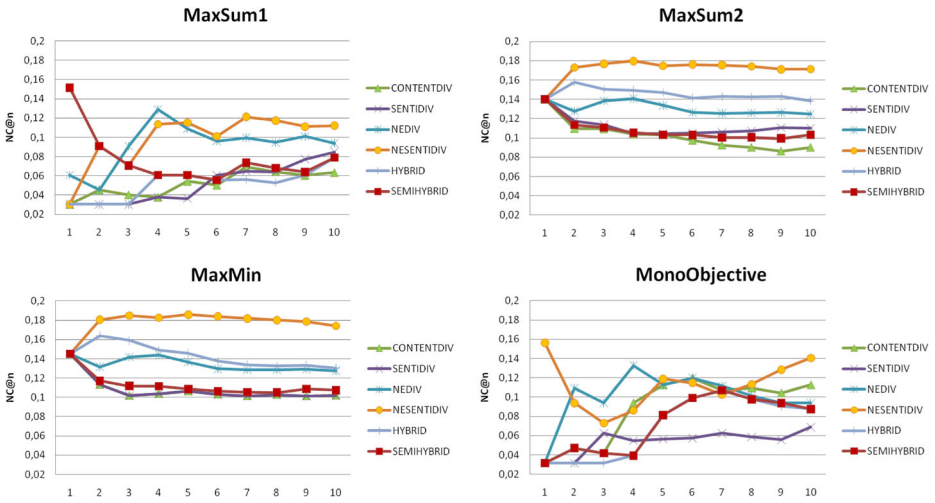

Fig. 6 Nugget coverage per algorithm

normalized, by definition, in the interval [0, 1]. The graphs are presented per algorithm, for clarity of presentation. For Nugget Coverage, the first observation is that, algorithms *MAXSUM2* and *MAXMIN* clearly outperform the other two ones, and, also, present a more consistent behavior, w.r.t. position for all tested variations. The second observation is that the baseline method (*CONTENTDIV*) is almost always outperformed by all variations of our approach, even when considering each algorithm separately. Finally, the variation of combining Named Entities and Sentiment (*NESENTIDIV*) distinctively outperforms all other variations (and of course the baseline), followed by the variation of combining all criteria (*HYBRID*) and considering only Named Entities (*NEDIV*).

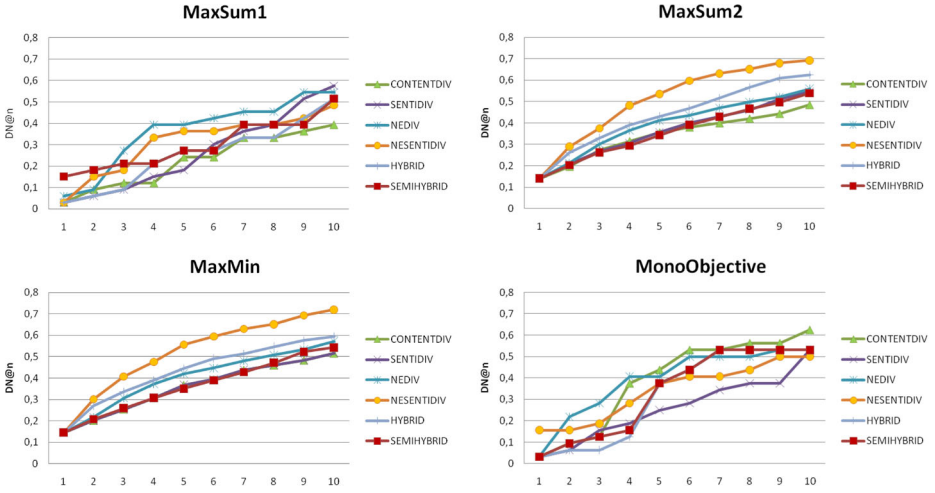

Fig. 7 Distinct nugget coverage per algorithm

Table 4 Nugget Coverage at position 5

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.055	0.036	0.109*	0.115*	0.061*	0.061	NESENTIDIV
MAXSUM2	0.103	0.104	0.134	0.175*	0.147*	0.103	NESENTIDIV
MAXMIN	0.106	0.123	0.137	**0.186***	0.146	0.108	NESENTIDIV
MONO	0.113	0.056	0.113	0.119*	0.081	0.081	NESENTIDIV
Best Algorithm per criterion	MONO	MAXMIN	MAXMIN	MAXMIN	MAXSUM2	MAXMIN	MAXMIN/ NESENTIDIV

Table 5 Nugget Coverage at position 10

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.064	0.085	0.094	0.112*	0.079*	0.079	NESENTIDIV
MAXSUM2	0.090	0.110	0.125	0.171*	0.139*	0.103	NESENTIDIV
MAXMIN	0.102	0.115	0.128	**0.174***	0.131	0.107	NESENTIDIV
MONO	0.113	0.069	0.094	0.141*	0.088	0.088	NESENTIDIV
Best Algorithm per criterion	MONO	MAXMIN	MAXMIN	MAXMIN	MAXSUM2	MAXMIN	MAXMIN/ NESENTIDIV

Table 6 Distinct Nugget Coverage at position 5

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.242	0.182	0.394	0.364*	0.273*	0.273	NEDIV
MAXSUM2	0.355	0.357	0.411	0.536*	0.431	0.342	NESENTIDIV
MAXMIN	0.368	0.399	0.421	**0.556***	0.445	0.351	NESENTIDIV
MONO	0.438	0.250	0.406	0.375	0.375	0.375	CONTENTDIV
Best Algorithm per criterion	MONO	MAXMIN	MAXMIN	MAXMIN	MAXMIN	MONO	MAXMIN/ NESENTIDIV

Table 7 Distinct Nugget Coverage at position 10

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.394	0.576	0.546	0.485*	0.515*	0.515	SENTIDIV
MAXSUM2	0.485	0.545	0.560	0.692*	0.625*	0.538	NESENTIDIV
MAXMIN	0.516	0.557	0.572	**0.720***	0.594	0.543	NESENTIDIV
MONO	0.625	0.531	0.531	0.500*	0.531	0.531	CONTENTDIV
Best Algorithm per criterion	MAXMIN	MAXMIN	MAXMIN	MAXMIN	MAXSUM2	MAXMIN	MAXMIN/ NESENTIDIV

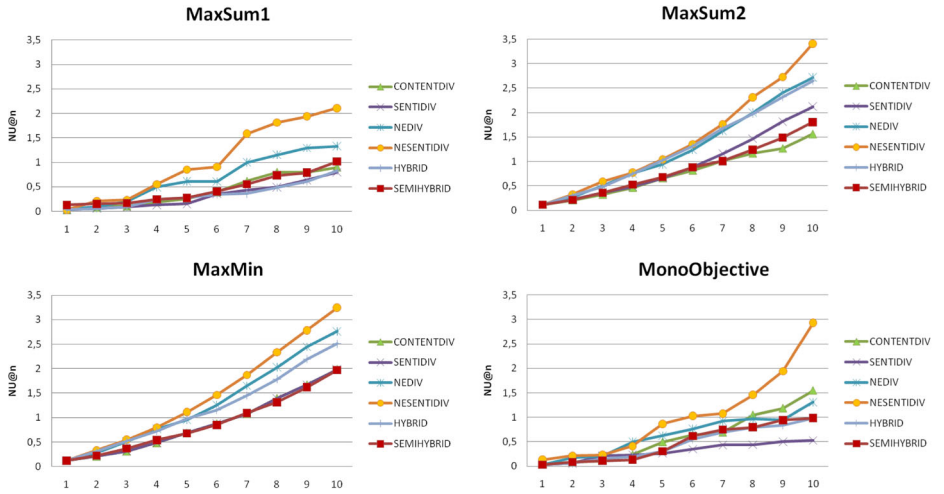

Fig. 8 Nugget uniformity per algorithm

For Distinct Nugget Coverage, more or less the same observations stand, with the exception of *CONTENTDIV* performing surprisingly good with *MONO-OBJECTIVE* algorithm, but, still, worse than *NESENTIDIV* in *MAXSUM2* and *MAXMIN*.

In order to better illustrate the quantitative differences in effectiveness, we consider, in Tables 4, 5, 6 and 7 the Nugget Coverage and Distinct Nugget Coverage values only for positions 5 and 10. We also mark the best performing algorithm per variation (last row), the best performing variation per algorithm (last column) and the overall best performing variation/algorithm combination. The combination *MAXMIN/NESENTIDIV* outperforms all the other combinations in all cases. Moreover, it increases the baseline performance by 65%, 54%, 27% and 15% for NC@5, NC@10, DN@5 and DN@10 respectively. The centesimal differences between *NESENTIDIV* and the baseline, for the above measures respectively are: 7.3%, 6.1%, 11.8% and 9.5%.

We also report the results on significance testing with T-Test[9] with confidence level 95%. At each row of Tables 4, 5, 6 and 7, significant scores compared to the *CONTENTDIV* baseline for the respective algorithm of the row are marked with asterisk. We note that the best performing (*MAXMIN/NESENTIDIV*) variation of our methods gives statistically significant scores compared to **all** *CONTENTDIV/algorithm* combinations.

Examining the above graphs and tables overall, first of all, it is clear that Named Entities is an important criterion to consider for diversifying user comments. The effectiveness of this criterion is even more boosted when being combined with Sentiment recognition around Named Entities. This is probably justified by the fact that it is expected that most topics described in news articles are somehow related to Persons or Organizations, so Named Entities help better capture these topics. On top of that, sentiment heterogeneity on these Named Entities, obviously, induces topic heterogeneity in comments. Second, it is demonstrated that more refined criteria than plain content diversity consistently perform better.

[9] http://www.socialresearchmethods.net/kb/stat_t.php

Table 8 Nugget Uniformity at position 5

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.259	0.149	0.612	0.850	0.272	0.272	SENTIDIV
MAXSUM2	0.659	0.662	0.940	1.041	1.018	0.678	CONTENTDIV
MAXMIN	0.688	0.793	0.954	1.112	0.974	0.677	SEMIHYBRID
MONO	0.496	0.265	0.621	0.866	0.304	0.304	SENTIDIV
Best Algorithm per criterion	MAXSUM1	MAXSUM1	MAXSUM1	MAXSUM1	MAXSUM1	MAXSUM1	MAXSUM1/ SENTIDIV

Table 9 Nugget Uniformity at position 10

Algorithm	CONTENTDIV	SENTIDIV	NEDIV	NESENTIDIV	HYBRID	SEMIHYBRID	Best Criterion per Algorithm
MAXSUM1	0.898	0.795	1.330	2.107	0.834	1.016	SENTIDIV
MAXSUM2	1.565	2.122	2.718	3.407	2.649	1.805	CONTENTDIV
MAXMIN	1.983	2.079	2.760	3.245	2.509	1.969	SEMIHYBRID
MONO	1.547	**0.5273**	1.309	2.929	0.984	0.984	SENTIDIV
Best Algorithm per criterion	MAXSUM1	MONO	MONO	MAXSUM1	MAXSUM1	MONO	MONO/ SENTIDIV

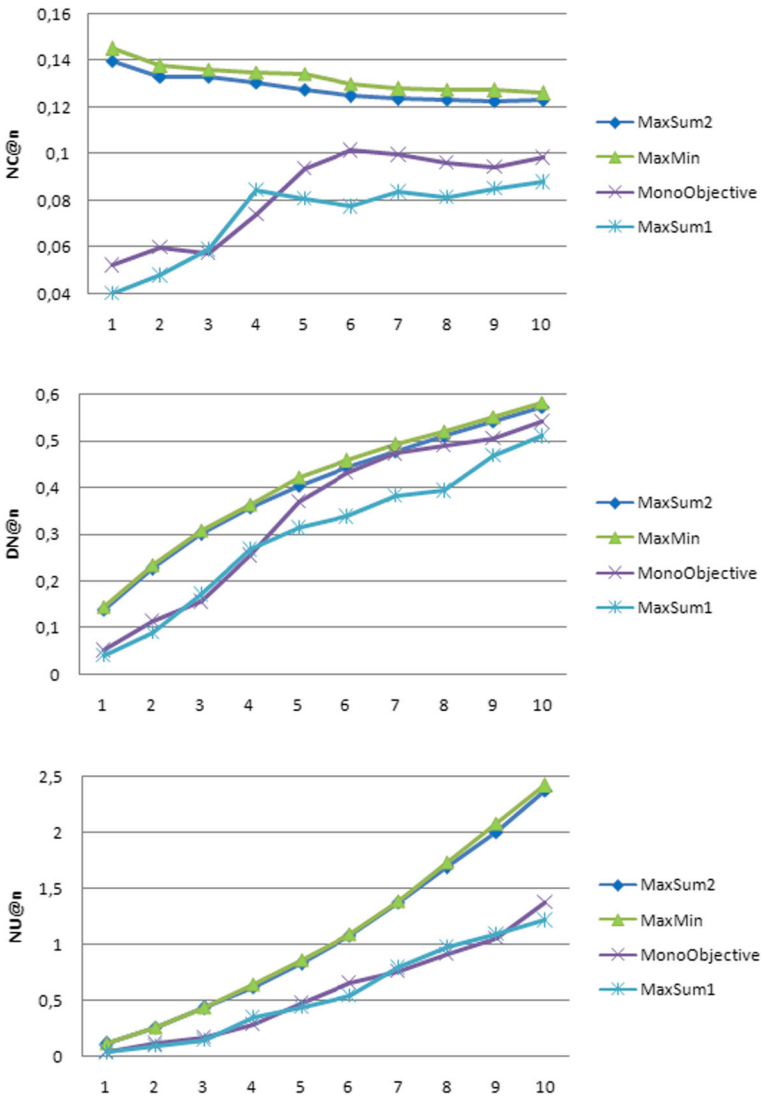

Fig. 9 Average effectiveness of each algorithm of all different criteria variations

Finally, *MAXMIN* is the best performing algorithm, followed by a slightly worse performing *MAXSUM2*. Both algorithms perform distinctively better that *MAXSUM1* and *MONO-OBJECTIVE*. The main difference between the two pairs of algorithms is that, *MAXMIN* and *MAXSUM2* compare, at each iteration, candidate comments with result comments, while the other two algorithms compare **only** candidate comments with each other.

Figure 8 and Tables 8, 9 illustrate the third measure, Nugget Uniformity, which measures, through a variance-like formula, the differences in nugget frequencies within each

method's result comment sets. We note that, in contrast with the previous ones, this measure is not normalized and lower value means better performance. Here, although the overall best performance for NU@5, NU@10 is achieved by *SENTIDIV*, there is not a variation that clearly outperforms all others. On top of that, it is obvious that most variation/algorithm combinations that perform good in the first two measures, perform relatively poor in the third one, and vice-versa. This can be justified by the fact that while the number of nuggets contained in the results of a method increases, it is expected that some nuggets that are more popular contribute more to the increase, while, when the overall number of nuggets is small, nugget frequency differences are expected to be small too. Moreover, it is expected that there exist some outlier nuggets, that is nuggets that appear only in very few comments, representing less important aspects of the article's topics. So, when the overall number of nuggets increases, these are expected to obtain low frequencies, affecting the Nugget Uniformity score of a method. Of course, Nugget Coverage and, especially Distinct Nugget Coverage are more important factors than Nugget Uniformity to take into account for the

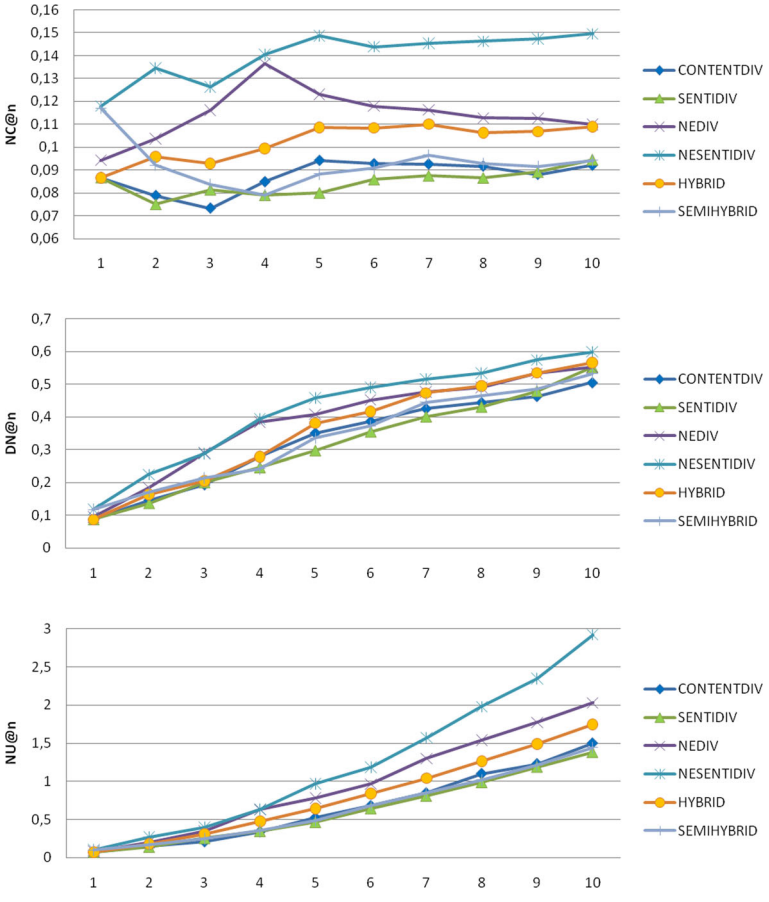

Fig. 10 Average effectiveness of each criterion variation of all different algorithms

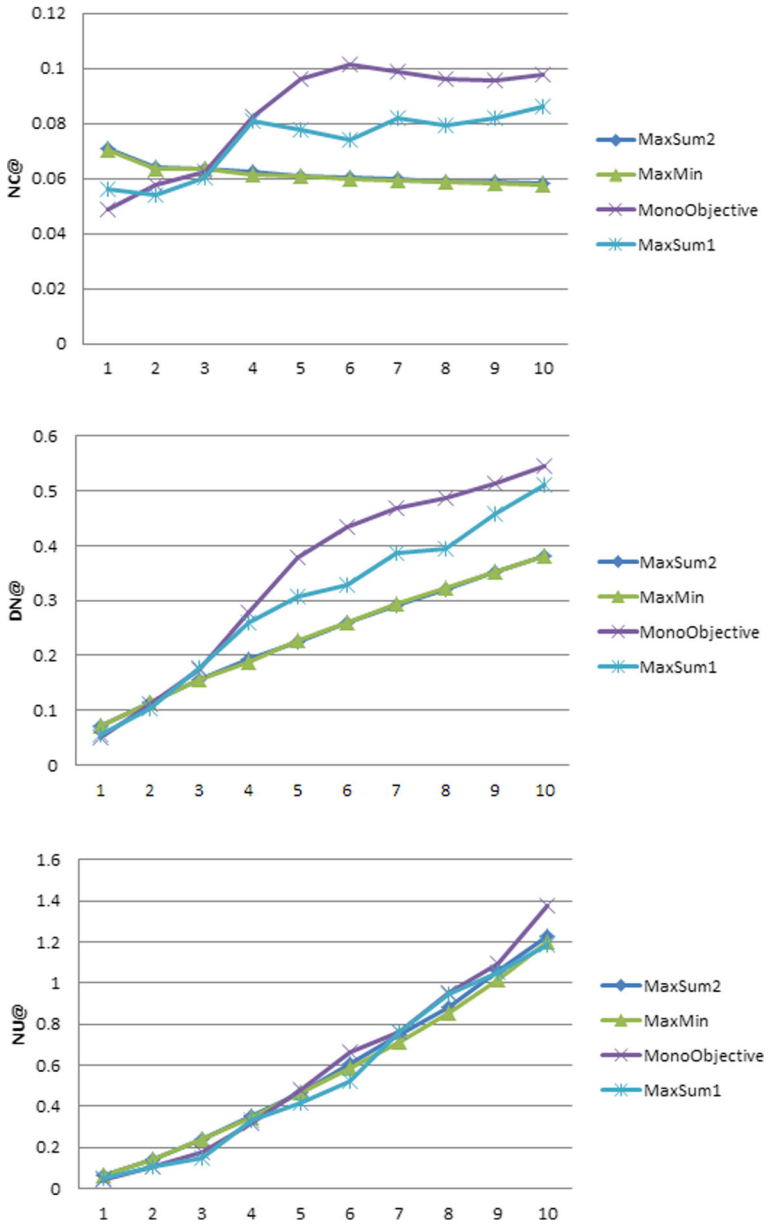

Fig. 11 Average effectiveness of each algorithm of all different criteria variations (automatically extracted nuggets)

task of diversification, so the ideal method should be selected based on them. On the other hand, there are combinations that comprise a middle ground, such as *NEDIV* and *HYBRID* variations in *MAXMIN* and *MAXSUM2* algorithms.

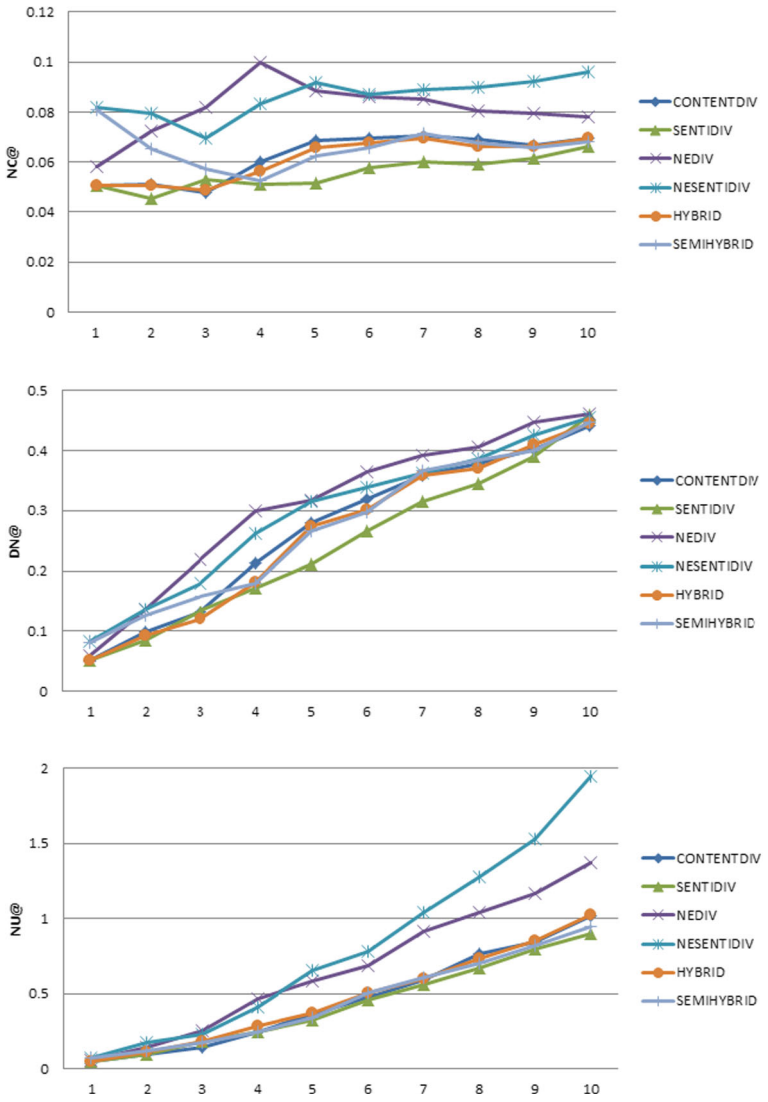

Fig. 12 Average effectiveness of each criterion variation of all different algorithms (automatically extracted nuggets)

We note that, contrary to the previous, precision like measures, we considered it meaningless to perform significance testing on Nugget Uniformity, since the measure is not normalized and the results would not have any intuitive importance.

In Fig. 9, the effectiveness of each algorithm averaged on all different variations is presented. The graph verifies the previous observations of *MAXMIN* and *MAXSUM2*

outperforming *MAXSUM1* and *MONO-OBJECTIVE* in Coverage measures and being outperformed in Uniformity measure.

Similarly, in Fig. 10, the effectiveness of each criterion variation averaged on all different algorithms is presented. The graph verifies the previous observations of *NESENTIDIV* significantly outperforming the rest variations in Coverage measures and *NEDIV* and *HYBRID* being suitable middle ground solutions, when all measures are considered important.

6.4.2 Measures on automatically extracted Information Nuggets

For simplicity, in what follows, we present the average effectiveness of each algorithm on all different criteria variations and the average effectiveness of each criterion variation on all different algorithms, respectively in Figs. 11 and 12.

The most importance difference, compared to the evaluation performed with manually extracted Nuggets, regards the average performance of each algorithm on all criteria, depicted in Fig. 11. Here, algorithms *MAXSUM1* and *MONO-OBJECTIVE* outperform *MAXMIN* and *MAXSUM2* in the measures of Nugget Coverage and Distinct Nugget Coverage, while all algorithms demonstrate similar performance on the measure of Nugget Uniformity.

On the other hand, when averaging the effectiveness of all algorithms over each criterion, the results are consistent with the results of Section 6.4.1: *NESENTIDIV* still performs better in terms of Coverage, however *NEDIV* has almost equal performance, while it performs better in Nugget Uniformity. Thus, it can be considered the best middle ground solution. Finally, we can see that on Distinct Nugget Coverage, all criteria end up performing almost the same at position 10 ($DN@10$), although still, *NEDIV* and *NESENTIDIV* perform slightly better. This can probably be attributed to the automatic topic extraction tools producing more rigid and overspecialized Nuggets, compared to the human extraction process, since they are based on the identified terms themselves and not the concepts they represent. This way, the automatic extraction process is expected to lead to Nuggets that do not allow implicit concepts to be identified through comments.

6.4.3 Discussion

The general conclusion of the two rounds of evaluation performed in Sections 6.4.1 and 6.4.2 are the following: (a) The effectiveness of each diversification criterion is not significantly influenced by the choice of manually or automatically extracted Information Nuggets as a measure of evaluation. That is, the criteria of sentiment around Named Entities and Named Entities constistently perform better than the rest ones (including the baseline criterion of content). (b) The relative effectiveness of diversification algorithms is affected by the choice of Nuggets, however, the standalone behavior of the graph of each algorithm remains the same. This can be probably attributed to some (important but implicit) manually extracted Information Nuggets, that are identified on specific comments, not being able to be recognized by automatic extraction tools and thus, decreasing the measurement values of an algorithm, but not significantly influencing the shape of their curves. (c) In both evaluation scenarios, the graphs of Distinct Nugget Coverage show that there is still room for improvement/refinement on the diversification criteria and algorithms: the maxi-

mum coverage of Distinct Nuggets achieved by any applied variation is 60%. This consists one of our main goals in our future steps and is closely related to examining new criteria (User Co-commenting Behavior) and refining old ones, as well as fine-tuning the applied diversification algorithms.

7 Conclusion

In this paper, we presented our approach on diversifying user comments on news articles. We introduced comment-specific diversification criteria and applied them on four heuristic diversification algorithms (three state of the art and one proposed by us algorithm variation), by defining the proper initialization and score aggregation functions. The experimental evaluation, based on adapted definition of information nuggets and setting-specific evaluation measures demonstrated the effectiveness of applying our proposed diversification criteria, as opposed to applying plain content diversity on news articles comments. Finally, the implemented framework is general enough to be adapted to other settings, such as forum discussions, tweets, blog posts and comments, etc. We implemented the above methods into an initial, prototype system that we intend to extend into a web application that will be able to produce diverse set of comments on news portal, forums, etc.

Our future work lies on enhancing the User Co-commenting Behavior criterion by applying topical clustering techniques in order to reduce the sparseness of the respective feature vectors and test its effectiveness. We also intend to test whether hybrid algorithm combinations could achieve even better effectiveness. Further, we plan to examine the effectiveness of our framework in other settings, e.g. tweets around topics, and identify possibly required improvements/adaptations so that the framework better functions in the respective settings. Also, an important direction is exploiting threads of comments (comment replies to previous comments) to refine the diversification process. Finally, we intend to further investigate the impact of several criteria (e.g. which NE categories affect diversification most, or try different sentiment/NE extraction tools), as well as to test more topic (Information Nuggets) extraction tools and compare the outcomes to our current evaluation results.

Acknowledgments This research is conducted as part of the EU project ARCOMEM[10] FP7-ICT-270239.

Appendix

Information nuggets of evaluated articles

[10] http://www.arcomem.eu/

Table 10 Evaluation articles abstracts and corresponding nuggets

Article's main topic	US consumption rate
Article's abstract	Jared Diamond Op-Ed article holds that biggest global concern is resource consumption rate; notes that consumption rate in North America is 32 times higher than in developing world; says US promise that any country that adopts free-market economy can enjoy 'first-world lifestyle' is cruel hoax; contends that if China's per capita consumption rates rise to US levels, world will run out of resources at even faster rate; says it is futile to tell other countries not to reach for consumption rate that Americans already enjoy; contends that present rate of US consumption is unsustainable; says American consumption is wasteful and contributes little or nothing to quality of life; says that US consumption rates could be lowered if there was political will to tackle problem; drawing
Article's Information Nuggets	Consumption rate, Freemarket economy/lifestyle, China consumption, US consumption, Developing world consumption, Political will to solve the problem, Consumption comparisons
Article's main topic	Democrats nominations
Article's abstract	Frank Rich Op-Ed column contends that if Hillary Clinton wins Democratic presidential nomination, Republicans will enjoy having both Bill Clinton and Hillary as targets; says now that Bill is ubiquitous on campaign trail, both his past and his post-presidency will be vetted; suggests that John McCain would be strong candidate against Hillary; says Barack Obama represents change and would have stronger argument against McCain; IN the wake of George W. Bush, even a miracle might not be enough for the Republicans to hold on to the White House in 2008. But what about two miracles? The new year's twin resurrections of Bill Clinton and John McCain, should they not evaporate, at last give the

Table 10 (continued)

	G.O.P. a highly plausible route to victory. Amazingly, neither party seems to fully. Any Democrat who seriously thinks that Bill Clinton will fade away if Hillary wins the party nomination is a Democrat who, as the man said, believes in fairy tales.
Article's Information Nuggets	Democratic presidential nomination, Republicans, Clinton unite Republicans, Bill Clinton vetted, Hillary Clinton influenced by Bill Clinton, John McCain, Barack Obama change, Barack Obama stronger than John McCain, George W. Bush, Hillary will work with Bill
Article's main topic	Prescriptions decrease: consequences/reasons
Article's abstract	Research firm IMS Health says number of prescriptions dispensed through August was lower than in first eight months of last year, evidence that people are scaling back on medications in effort to save money in economic hard times; doctors report that they have patients who have stopped taking important drugs; trend, if it continues, could have profound implications and could eventually raise nation's total health care bill and lower standard of living; many people do not stop taking drugs, rather they split their pills to extend length of time before they have to refill prescription; overall spending on medicines is still highest in world, estimated at 286.5 billion in 2007; other factors that could be causing decline include higher co-payments
Article's Information Nuggets	Prescriptions decrease, Decrease to save money, Decrease to important drugs, Might raise eventual cost, Lowers standard of living, Split pills, Higher copayments, Chose gas and meal over drugs
Article's main topic	Obama measures on financial crisis

Table 10 (continued)

Article's abstract	David Brooks Op-Ed column warns that steps taken to relieve financial crisis tend to punish responsible people along with the profligate; says fixes put in place by Obama economic team are not integrated response, although they appear to be driven by sense of moderation and restraint; concludes that if those who have been greedy are not saved, then entire economy will not be stabilized
Article's Information Nuggets	Measures punish responsible, Measures punish irresponsible, Measures are moderate, Measures are not integrated, Save greedy to stabilize economy, People should face consequences/justice, Economy is interwoven

Table 11 Evaluation articles abstracts and corresponding nuggets

Article's main topic	Relation between successful people/elite colleges
Article's abstract	David Leonhardt Economic Scene column on contends student bodies at top colleges which, despite other ways they have become diverse, are still overwhelmingly affluent; says system does not serve national interest, but rather the interests of wealthy individuals; The last four presidents of the United States each attended a highly selective college. All nine Supreme Court justices did, too, as did the chief executives of General Electric (Dartmouth), Goldman Sachs (Harvard), Wal-Mart (Georgia Tech), Exxon Mobil (Texas) and Google (Michigan). Like it or not, these colleges have outsize influence on American
Article's Information Nuggets	Most top college students are rich, Serves no national interest, Serves interests of the rich, Presidents/Judges/Corporates from high colleges, Admission should consider national interest
Article's main topic	Tax evasion
Article's abstract	75,000 IN INCOME WAS UNREPORTED ON RANGEL TAXES Representative Charles B. Rangel has earned more than 75,000 in rental income from a villa he has owned in the Dominican Republic since 1988, but never reported it on his federal or state tax returns, according to a lawyer for the congressman and documents from the resort. Mr. Rangel, chairman of the House Ways and Means Committee, which writes Rep Charles B Rangel has earned more than 75,000 in rental income from Dominican Republic villa he has owned since 1988, but never reported it on his tax returns; lawyer for Rangel says he will most likely file amendments to his tax returns for years in question; says Rangel will probably have no federal tax liability because he considered property investment rather than

Table 11 (continued)

	vacation home and is therefore entitled to deduct depreciation on property; says Rangel did not realize he had to declare money as income
Article's Information Nuggets	Tax evasion, Charles B. Rangel, Tax, Finance
Article's main topic	Obama's anti-foreclosure plan
Article's abstract	Helping the House Poor, President Obama's anti-foreclosure plan, which took effect on Wednesday, is better than anything attempted by the Bush administration, but it is at best one step forward and, unfortunately, may prove to be fundamentally flawed. The program's success ultimately will rest on whether the administration is willing to intensify its efforts and, as, Editorial asserts that while Pres Obama's anti-foreclosure plan is positive step, it may prove to be fundamentally flawed; contends that emphasis on lowering interest rates may make monthly payments more affordable for homeowners, but it does not change that they still owe more than property is worth; argues that more effective course is to focus on reducing principal loan balance to rebuild home equity and give borrowers incentive to avoid defaulting
Article's Information Nuggets	Mortgage, Housing bubble, Foreclosure, Real estate, Obama, Politics, Debt settlement
Article's main topic	Scandal with politician and bankers
Article's abstract	A Representative, Her Ties and a Bank Meeting Top federal regulators say they were taken aback when they learned that a California congresswoman who helped set up a meeting with bankers last year had family financial ties to a bank whose chief executive asked them for up to 50 million in special bailout funds. Representative Maxine Waters, Democrat of California, requested the September Rep Maxine Waters

Table 11 (continued)

	asked top federal regulators to meet in September 2008 with executives of minority-owned banks hurt by federal takeover of Fannie Mae and Freddie Mac without telling them that her husband, Sidney Williams, had served on board of one of largest of those banks, OneUnited, until early in 2008 and has owned at least 250,000 of its stock; meeting was intended to address losses suffered by banks as whole, but Kevin Cohee, One United's chief executive, made special plea at meeting for 50 million in cash for his bank; bank did not receive that money, but did receive 12 million in December through Treasury's Troubled Asset Relief Program; OneUnited has been faulted by regulators for not lending enough money to low-income residents in Miami and for providing Cohee with 6.4 million beachfront compound in Santa Monica, Calif; bank's critics say episode shows how special access arranged through lawmaker with financial ties to bank compromised integrity of TARP effort; photos
Article's Information Nuggets	OneUnited Bank, Maxine Waters, Bailout, Troubled Asset Relief Program, Economics, Politics, Financial Institution, Kevin Coheet

Table 12 Evaluation articles abstracts and corresponding nuggets

Article's main topic	Financial reform related to elections
Article's abstract	Wall Street Casino Congressional Republicans have concluded that screaming foul about the banking bailout and blocking financial reform is a clever strategy for the fall elections. This approach ignores some pretty basic history: that the banks imploded while Republicans held Congress and the White House; that President George W. Bush started the rescue; that many Editorial scores banks like Goldman Sachs, which turned American financial system into casino by selling incomprehensible mortgage-backed products while placing bets against them; charges that society bore cost of products that packed enormous capacity for economic destruction; says information coming out of Senate hearings on role of Wall Street in economic crisis should make it clear that bill instituting financial reform should be passed
Article's Information Nuggets	Subprime mortgage crisis, Financial Institution, Politics, Wall Street, Goldman Sachs, Bailout, Regulation
Article's main topic	Federal loans on energy programs
Article's abstract	Hooray for Federal Loans! In the firestorm over Solyndra, three main criticisms have emerged. The first is that Solyndra wasn't ready for prime time and that the Department of Energy, which gave it a 535 million federally guaranteed loan, should have known as much. The second is that Solyndra used political influence to land a loan that was destined to blow up. And the Joe Nocera Op-Ed column argues that Department of Energy program, which offers federally guaranteed loans to alternative energy companies, has done more good than bad, despite criticism over funding of Solyndra, solar energy company that went bankrupt; notes that federal government, not the private sector, is willing to invest in alternative energies during difficult financial times
Article's Information Nuggets	Solyndra, Loan, Politics, Finance, Alternative energy, Solar energy, Bankrupt

References

Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S. (2009). Diversifying search results. In *Proceedings of the second international conference on web search and web data mining* (WSDM 2009) (pp.5-14).

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (SIGIR '98)(pp.335-336).

Chandra, B., & Halldórsson, M.M. (2001). Approximation algorithms for dispersion problems. *Journal of Algorithms*, *38*(2), 438–465.

Chen, H., & Karger, D.R. (2006). Less is more: Probabilistic models for retrieving fewer relevant documents. In: *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (SIGIR '06)(pp. 429-436).

Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation.In: P*roceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (SIGIR '08)(pp. 659–666).

Diakopoulos, N., & Naaman, M. (2011). Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (CSCW '11)(pp. 133–142).

Drosou, M., & Pitoura, E. (2010). Search result diversification. *ACM SIGMOD record*, *39*(1), 41–47.

Erkut, E. (1990). The discrete p-dispersion problem. *Operations Research Letters*, *46*(1), 48–60.

Erkut, E., Ülküsal, Y., Yeniçerioglu, O. (1994). A comparison of p-dispersion heuristics. *Computers Operations Research*, *21*(10), 1103–1113.

Finkel, J.R., Grenager, T., Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: *Proceedings of the 43nd annual meeting of the association for computational linguistics* (ACL '05)(pp. 363–370).

Giannopoulos, G., Weber, I., Jaimes, A., Sellis, T. (2012). Diversifying User Comments on News Articles. In: *Proceedings of the 13th international conference web information systems engineering* (WISE '12)(pp. 100–113).

Gollapudi, S., & Sharma, A. (2009). An axiomatic approach for result diversification. In: *Proceedings of the 18th international conference on World wide web* (WWW '09)(pp. 381–390).

Hassin, R., Rubinstein, S., Tamir, A. (1997). Approximation algorithms for maximum dispersion. *Operations Research Letters*, *21*(3), 133–137.

Herring, S.C., Kouper, I., Paolillo, J.C., Scheidt, L.A., Tyworth, M., Welsch, P., Wright, E., Ning, Y. (2005). Conversations in the blogosphere: an analysis "from the bottom up". In: *Proceedings of the 38th annual hawaii international conference on system sciences*, (HICSS '05)(pp. 107b–107b).

Hu, M., Sun, A., Lim, E. (2008). Comments-oriented document summarization: Understanding documents with readers' feedback. In: *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (SIGIR '08)(pp. 291–298).

Kucuktunc, O., Cambazoglu, B.B., Weber, I., Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for Yahoo! answers. In: *Proceedings of the 5th ACM international conference on Web search and data mining* (WSDM'12)(pp. 633–642).

Mishne, G.A., & Glance, N. (2006). Leave a Reply: An analysis of weblog comments. In: *Proceedings of the WWW 2006 workshop on weblogging ecosystem: aggregation, analysis and dynamics, at WWW ': the 15th international conference on world wide web*.

Munson, S.A., & Resnick, P. (2010). Presenting diverse political opinions: How and how much. In: *Proceedings of the 28th international conference on Human factors in computing systems* (CHI '10)(pp. 1457–1466).

Park, S., Ko, M., Kim, J., Liu, Y., Song, J. (2011). The politics of comments: predicting political orientation of news stories with commenters sentiment patterns. In: *Proceedings of the ACM 2011 conference on computer supported cooperative work* (CSCW '11)(pp. 113–122).

Potthast, M. (2009). Measuring the descriptiveness of web comments. In: *Proceedings of the 32nd international ACM SIGIR conference on research and development* (SIGIR '09)(pp. 724–725).

Ravi, S.S., Rosenkrantzt, D.J., Tayi, G.K. (2007). Approximation algorithms for facility dispersion. In Gonzalez, T.F. (Ed.) *Handbook of Approximation algorithms and metaheuristics*: Chapman & Hall/CRC.

Shmueli, E., Kagian, A., Koren, Y., Lempel, R. (2012). Care to Comment?Recommendations for Commenting on News Stories. In: *Proceedings of the 18th international conference on World wide web WWW '12*, to appear.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558.

Tsagkias, E., Weerkamp, W., de Rijke, M. (2009). Predicting the volume of comments on online news stories. In: *Proceedings of the 18th ACM conference on Information and knowledge management* (CIKM '09)(pp.1765–1768).

Tsagkias, E., Weerkamp, W., de Rijke, M. (2010). News Comments: exploring, modeling, and online predicting. In: *Proceedings of the 32nd european conference on information retrieval* (ECIR '10)(pp. 109–203).

Vallet, D., & Castells, P. (2012). Personalized diversification of search results. In: *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (SIGIR '12)(pp. 841–850).

Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., Yahia, S.A. (2008). Efficient computation of diverse query results. In: *Proceedings of the 2008 IEEE 24th international conference on data engineering* (ICDE '08)(pp. 228–236).

Li, Q., Wang, J., Chen, Y.P., Lin, Z. (2010). User comments for news recommendation in forum-based social media. *Information Sciences: An International Journal*, *180*(24), 4929–4939.

Wong, D., Faridani, S., Bitton, E., Hartmann, B., Goldberg, K. (2011). The diversity donut: enabling participant control over the diversity of recommended responses. In: *Proceedings of the 2011 annual conference extended abstracts on human factors in computing systems* (CHI EA '11)(pp. 1471–1476).