261 before passing the classified documents of each cluster 255a, 255b, 255c to the predicted documents 212.

[0035] Text summarization has emerged as an important research area in recent years. A lot of keyword extraction approaches are widely used to provide a compact representation of document content. However, not all existing keyword/keyphrase extraction techniques could be applied to a group of documents to differentiate it from the others. The idea of reducing important information present in a group of similar documents makes clustering/classification solutions easy to understand and helps with the query optimization.

[0036] For keyword extraction, a new approach, TF-IDF, is introduced which normalizes keyword weights with a length of the documents, the frequency of a term across the documents, and a maximum frequency of the term in a given group of documents. The well-known TF-IDF approach has an issue when applying it to a set of the documents: it is very sensitive to document length and term frequency. For example, when one term appears in all the documents in a cluster a hundred times altogether and another word appears a hundred times in only one document, they will get the same weight in representation of the current cluster. The goal is to find the proper normalization techniques to include document engagement into the weight calculation process.

[0037] On the first normalization approach term frequency is multiplied by percentage of the documents containing these terms as shown in Equations (1) and (2).

$$TDF = TF \times Doc\% \quad (1)$$

$$T_{norm} = tfidf_{i,j} \times TDF \times \log\left(\frac{N}{df_i}\right)$$

wherein $tf_{i,j}$ is the total number of occurrences of i in j,
df_i is the total number of documents (speeches) containing i,
N is the total number of documents (speeches) (2)

[0038] The term frequency is calculated in a cluster/class and the term frequency is multiplied into a percentage of the documents where a word appears. Then, the IDF score is calculated for the words per cluster. On a second normalization approach, the already normalized term frequency is multiplied by a percentage of the documents containing these terms as shown in Equations (3) and (4).

$$TNDF = T_{norm} \times Doc\% \quad (3)$$

$$DRAKE = tfidf_{i,j} \times TNDF \times \log\left(\frac{N}{df_i}\right)$$

wherein $tf_{i,j}$ is the total number of occurrences of i in j,
df_i is the total number of documents (speeches) containing i,
N is the total number of documents (speeches) (4)

[0039] The term frequency is calculated in a cluster/class and the document frequency (i.e., how many documents include the term) is calculated. Then, the IDF score is calculated for the words per cluster.

[0040] Rapid automatic keyword extraction (RAKE) is a keyword extraction method that is extremely efficient, operates on individual documents to enable application to dynamic collections, is easily applied to new domains, and operates well on multiple types of documents, particularly those that do not follow

CLAIMS

What is claimed is:

1. A computer-implemented method for data loss prevention comprising:

offline unsupervised text summarization of specific groups of documents comprising:

keyword/keyphrase extraction from the specific groups of documents; and

normalization of candidate keywords/keyphrases;

vectorization of the extracted keywords/keyphrases from the specific groups of documents; and

quantitative performance measurement of the keyword/keyphrase extraction to derive keywords and/or keyphrases suitable for data loss prevention.

2. The method in accordance with Claim 1 further comprising autolabelling categories and confidential statuses of documents of the specific groups of documents in response to the keyword/keyphrase extraction.

3. The method in accordance with Claim 1 wherein the combination of keywords and/or key phrases comprises both positive and negative combinations.

4. The method in accordance with Claim 1 wherein the keyword/keyphrase extraction comprises keyword extraction by TDF-IDF.

5. The method in accordance with Claim 1 or Claim 4 wherein the keyword/keyphrase extraction comprises keyphrase extraction by DRAKE.

6. The method in accordance with Claim 1 or Claim 4 or Claim 5 wherein the keyword/keyphrase extraction comprises keyphrase extraction by EmbedDocRank.

7. The method in accordance with Claim 1 wherein the vectorization of the extracted keywords/keyphrases from the specific groups of documents comprises valuation based on information gain and cross-validation.