Privacy-Preserving Important Passage Retrieval

Luís Marujo[1,2,3], José Portêlo[2,3], David Martins de Matos[2,3], João P. Neto[2,3], Anatole Gershman[1], Jaime Carbonell[1], Isabel Trancoso[2,3], Bhiksha Raj[1]
[1] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Instituto Superior Técnico, Lisboa, Portugal;
[3] INESC-ID, Lisboa, Portugal
{lmarujo,anatoleg,jgc,bhiksha}@cs.cmu.edu,
{jose.portelo,david.matos,joao.neto,isabel.trancoso}@inesc-id.pt

ABSTRACT

State-of-the-art important passage retrieval methods obtain very good results, but do not take into account privacy issues. In this paper, we present a privacy preserving method that relies on creating secure representations of documents. Our approach allows for third parties to retrieve important passages from documents without learning anything regarding their content. We use a hashing scheme known as Secure Binary Embeddings to convert a key phrase and bag-of-words representation to bit strings in a way that allows the computation of approximate distances, instead of exact ones. Experiments show that our secure system yield similar results to its non-private counterpart on both clean text and noisy speech recognized text.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]; I.2.7 [**Natural Language Processing**]: Text analysis; K.4.1 [**Computers and Society**]: Public Policy Issues—*privacy*

General Terms

Algorithms, Experimentation

Keywords

Secure Passage Retrieval, Important Passage Retrieval, KP-Centrality, Secure Binary Embeddings, Data Privacy, Automatic Key Phrase Extraction

1. INTRODUCTION

Important Passage Retrieval (IPR) is the problem of extracting the most important passages in a body of text. By "important", we mean those passages that capture most of the key information the text is attempting to convey. Of the various solutions proposed, state-of-the-art solutions for IPR based on *centrality* achieve excellent results [24].

PIR'14,Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security, SIGIR 2014 Workshop July 11th, 2014, Gold Coast, Australia

A potential problem to the deployment of such methods is that they usually assume that the input data are of public domain. However, this data may come from social network profiles, medical records or other private documents, and their owners may not want to, or even be allowed to share it with third parties. Consider the scenario where a company has millions of classified documents. The company needs to retrieve the most important passages from those documents, but lacks the computational power or know-how to do so. At the same time, they can not give access to the documents to a third party with such capabilities because they may contain sensitive information. As a result, the company must *obfuscate* their own data before sending it to the third party, a requirement that is seemingly at odds with the objective of extracting important passages from it.

In this paper, we propose a new *privacy-preserving* technique for IPR based on Secure Binary Embeddings (SBE) [3] that enables exactly this – it provides a mechanism for obfuscating the data, while still achieving near state-of-the-art performance in IPR.

SBEs are a form of locality-sensitive hashing which convert data arrays such as bag-of-words vectors to obfuscated bit strings through a combination of random projections followed by banded quantization. The method has information theoretic guarantees of security, ensuring that the original data cannot be recovered from the bit strings. At the same time, they also provide a mechanism for *locally* computing distances between vectors that are close to one another without revealing the global geometry of the data, consequently enabling tasks such as IPR. This is possible because, unlike other hashing methods which require exact matches for performing classification tasks, SBEs allows for a near-exact matching: the hashes can be used to estimate the distances between vectors that are very close, but provably provide no information whatsoever about the distance between vectors that are farther apart. The usefulness of SBE has already been shown for implementing a privacy-preserving speaker verification system [21] yielding promising results.

The remainder of the paper is structured as follows. In Section 2 we briefly present some related work regarding Important Passage Retrieval and privacy-preserving methods in IR. In Section 3 we detail the two stages of the important passage retrieval technique. Section 4 presents the method for obtaining a secure representation method. We describe our approach to privacy-preserving important passage retrieval in Section 5. Section 6 describes the dataset used and illustrates our approach with some experiments. Finally, we present some conclusions and plans for future work.

2. RELATED WORK

2.1 Important Passage Retrieval

Text and speech information sources influence the complexity of the important passage retrieval approaches differently. For textual passage retrieval, it is common to use complex information, such as syntactic [30], semantic [28], and discourse information [29], either to assess relevance or reduce the length of the output. However, speech important passage retrieval approaches have an extra layer of complexity, caused by speech-related issues like recognition errors or disfluencies. As a result, it is useful to use speech-specific information (e.g.: acoustic/prosodic features [15], recognition confidence scores [33]), or by improving both the assessment of relevance and the intelligibility of automatic speech recognizer transcriptions (by using related information [23]). These problems not only increase the difficulty in determining the salient information, but also constrain the applicability of passage retrieval techniques to speech passage retrieval. Nevertheless, shallow text summarization approaches such as Latent Semantic Analysis (LSA) [9] and Maximal Marginal Relevance (MMR) [4] seem to achieve performances comparable to the ones using specific speech-related features [20]. In addition, discourse features start to gain some importance in speech retrieval [15, 35].

Closely related to the important passage retrieval used by this work are approaches using the unsupervised key phrase extraction methods. These methods are used to reinforce passage retrieval [34, 31, 11, 25, 27]. Namely, they propose the use of key phrases to summarize news articles [11] and meetings [25]. In [11], the authors explored both supervised and unsupervised methods with a limited set of features to extract key phrases as a first step towards important passage retrieval. Furthermore, the important passage retrieval used in this work adapts the centrality retrieval model, which plays an important role in the whole process. This kind of model adaptation is explored in [25], where the first stage of their method consists in a simple key phrase extraction step, based on part-of-speech patterns; then, these key phrases are used to define the relevance and redundancy components of a MMR summarization model.

Most of the IPR methods could be easily adapted to be secure using the method described in Section 4. We opted to use the KP-Centrality method described in the next section because it has the current state-of-the-art IPR method.

2.2 Privacy Preserving Methods

In this work, we focused on creating a method to perform important passage retrieval keeping the information in the original documents private. There is a large body of literature on important passage retrieval and privacy preserving or secure methods. To the best of our knowledge, the combination of both research lines has not been explored yet. However, there are some recent works combining information retrieval and privacy. Most of these works use data encryption [8, 17, 7, 12] to transfer the data in a secure way. This does not solve our problem because the content of the document would be decrypted by the retrieval method and therefore it would not remain confidential to the retrieval method. Another alternative secure information retrieval methodology is to obfuscate queries, which hides user topical intention [19], but does not secure documents content.

In many areas the interest in privacy-preserving methods where two or more parties are involved and they wish to jointly perform a given operation without disclosing their private information is not new, and several techniques such as Garbled Circuits [32], Homomorphic Encryption [18], Locality-Sensitive Hashing [6] have been introduced. However, they all have limitations regarding the Important Passage Retrieval task we wish to address. Until recently Garbled Circuits were extremely inefficient to use due to several intrinsic issues, and even now it is difficult to adapt them when the computation of non-linear operations is required. Solutions to many of these problems have been developed, such as performing offline computation of the oblivious transfers, using shorter ciphers, evaluating XOR gates for 'free', etc. [1]. Systems based on Homomorphic Encryption techniques introduce substantial amounts of computational overhead and usually require extremely long amounts of time to evaluate any function of interest. The Locality-Sensitive Hashing technique allows for near-exact match detection between data points, but does not provide any actual notion of distance, leading to degradation of performance in some applications. As a result, we decided to consider Secure Binary Embeddings [3] as the data privacy method for our approach, as it does not show any of the disadvantages mentioned above for the task at hand. We describe this technique in depth in Section 4.

3. IMPORTANT PASSAGE RETRIEVAL

To determine the most important sentences of an information source, we used the KP-Centrality model [24]. We chose this model for its adaptability to different types of information sources (e.g., text, audio and video) and state-of-the-art performance. It is based on the notion of combining key phrases with support sets. A support set is a group of the most semantically related passages. These semantic passages are selected using heuristics based on the passage order method [22]. This type of heuristic explore the structure of the input source to partition the candidate passages to be included in the support set in two subsets: the ones closer to the passage associated with the support set under construction and the ones further apart. These heuristics use a permutation, $d^i_1, d^i_2, \cdots, d^i_{N-1}$, of the distances of the passages s_k to the passage p_i, related to the support set under construction, with $d^i_k = dist(s_k, p_i)$, $1 \leq k \leq N-1$, where N is the number of passages, corresponding to the order of occurrence of passages s_k in the input source. The metric that is normally used is the cosine distance.

The KP-Centrality method consists of two steps, as illustrated in Figure 1. First, it extracts key phrases using a supervised approach [13] and combines them with a bag-of-words model in a compact matrix representation, given by:

$$\begin{bmatrix} w(t_1,p_1) & \dots & w(t_1,p_N) & w(t_1,k_1) & \dots & w(t_1,k_M) \\ \vdots & & & & & \vdots \\ w(t_T,p_1) & \dots & w(t_T,p_N) & w(t_T,k_1) & \dots & w(t_T,k_M) \end{bmatrix}, \quad (1)$$

where w is a function of the number of occurrences of term t_i in passage p_j or key phrase k_l, T is the number of terms and M is the number of key phrases. Then, using a segmented information source $I \triangleq p_1, p_2, \dots, p_N$, a support set S_i is computed for each passage p_i using:

$$S_i \triangleq \{s \in I \cup K : sim(s, p_i) > \varepsilon_i \wedge s \neq p_i\}, \quad (2)$$

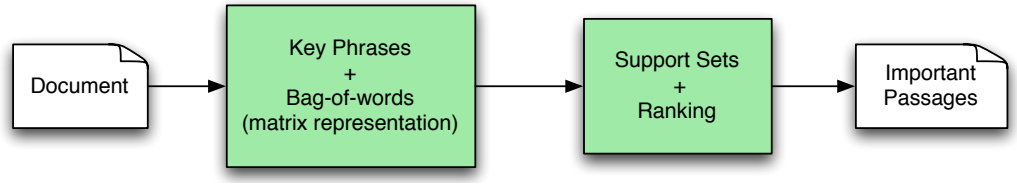

Figure 1: Flowchart of the Important Passage Retrieval method.

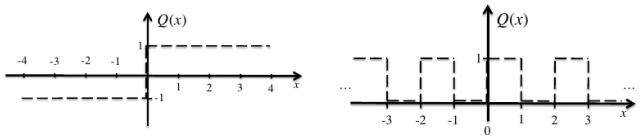

Figure 2: 1-bit quantization functions.

for $i = 1, \ldots, N+M$. Passages are ranked excluding the key phrases K (*artificial passages*) according to:

$$\underset{s \in (\cup_{i=1}^{n} S_i) - K}{\arg\max} \left| \{S_i : s \in S_i\} \right|. \quad (3)$$

4. SECURE BINARY EMBEDDINGS

A Secure Binary Embedding (SBE) is a scheme that converts real-valued vectors to bit sequences using band-quantized random projections. These bit sequences, which we will refer to as *hashes*, possess an interesting property: if the Euclidean distance between two vectors is lower than a threshold, then the Hamming distance between their hashes is proportional to the Euclidean distance between the vectors; if it is higher, then the hashes provide no information about the true distance between the two vectors. This scheme relies on the concept of Universal Quantization [2], which redefines scalar quantization by forcing the quantization function to have non-contiguous quantization regions.

Given an L-dimensional vector $\mathbf{x} \in \mathbb{R}^L$, the universal quantization process converts it to an M-bit binary sequence, where the m-th bit is given by

$$q_m(\mathbf{x}) = Q\left(\frac{\langle \mathbf{x}, \mathbf{a}_m \rangle + w_m}{\Delta}\right). \quad (4)$$

Here $\langle , \rangle$ represents a dot product. $\mathbf{a}_m \in \mathbb{R}^L$ is a projection vector comprising L i.i.d. samples drawn from $\mathcal{N}(\mu = 0, \sigma^2)$, Δ is a precision parameter, and w_m is a random dither drawn from a uniform distribution over $[0, \Delta]$. $Q(\cdot)$ is a quantization function given by $Q(x) = \lfloor x \bmod 2 \rfloor$. We can represent the complete quantization into M bits compactly in vector form:

$$\mathbf{q}(\mathbf{x}) = Q\left(\boldsymbol{\Delta}^{-1}(\mathbf{A}\mathbf{x} + \mathbf{w})\right), \quad (5)$$

where $\mathbf{q}(\mathbf{x})$ is an M-bit binary vector, which we will refer to as the *hash* of $\mathbf{x}$. $\mathbf{A} \in \mathbb{R}^{M \times L}$ is a matrix composed of the row vectors $\mathbf{a}_m$, $\boldsymbol{\Delta}$ is a diagonal matrix with entries Δ, and $\mathbf{w} \in \mathbb{R}^M$ is a vector composed from the dither values w_m.

The universal 1-bit quantizer of Equation 4 maps the real line onto 1/0 in a banded manner, where each band is Δ_m wide. Figure 2 compares conventional scalar 1-bit quantization (left panel) with the equivalent universal 1-bit quantization (right panel).

The binary hash generated by the Universal Quantizer of Equation 5 has the following properties [3]: the probability that the i^{th} bits, $q_i(\mathbf{x})$ and $q_i(\mathbf{x}')$ respectively, of hashes of two vectors $\mathbf{x}$ and $\mathbf{x}'$ are identical depends only on the Euclidean distance $d_E = \|\mathbf{x} - \mathbf{x}'\|$ between the vectors and not on their actual values. As a consequence, the following relationship can be shown [3]: given any two vectors $\mathbf{x}$ and $\mathbf{x}'$ with a Euclidean distance d_E, with probability at most $e^{-2t^2 M}$ the normalized (per-bit) Hamming distance $d_H(\mathbf{q}(\mathbf{x}), \mathbf{q}(\mathbf{x}'))$ between the hashes of $\mathbf{x}$ and $\mathbf{x}'$ is bounded by:

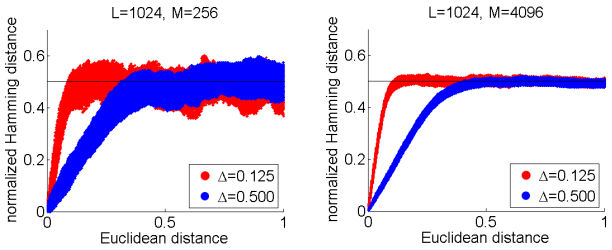

Figure 3: SBE behavior as a function of Δ, for two values of M.

$$\frac{1}{2} - \frac{1}{2} e^{-\left(\frac{\pi \sigma d_E}{\sqrt{2}\Delta}\right)^2} - t \leq d_H(\mathbf{q}(\mathbf{x}), \mathbf{q}(\mathbf{x}')) \leq \frac{1}{2} - \frac{4}{\pi^2} e^{-\left(\frac{\pi \sigma d_E}{\sqrt{2}\Delta}\right)^2} + t, \quad (6)$$

where t is the control factor. The above bound means that the Hamming distance $d_H(\mathbf{q}(\mathbf{x}), \mathbf{q}(\mathbf{x}'))$ is correlated to the Euclidean distance d_E between the two vectors, if d_E is lower than a threshold (which depends on Δ). Specifically, for small d_E, $E[d_H(\mathbf{q}(\mathbf{x}), \mathbf{q}(\mathbf{x}'))]$, the expected Hamming distance, can be shown to be bounded by $\sqrt{2\pi^{-1}}\sigma\Delta^{-1}d_E$, which is linear in d_E. However, if the distance between $\mathbf{x}$ and $\mathbf{x}'$ is higher than this threshold, then $d_H(\mathbf{q}(\mathbf{x}), \mathbf{q}(\mathbf{x}'))$ is bounded by $0.5 - 4\pi^{-2} exp\left(-0.5\pi^2\sigma^2\Delta^{-2}d_E^2\right)$, which rapidly converges to 0.5 and effectively gives us no information whatsoever about the true distance between $\mathbf{x}$ and $\mathbf{x}'$.

In order to illustrate how this scheme works, we randomly generated pairs of vectors in a high-dimensional space ($L = 1024$) and plotted the normalized Hamming distance between their hashes against the Euclidean distance between them (Figure 3). The number of bits in the hash is also shown in the figures.

We note that in all cases, once the normalized distance exceeds Δ, the Hamming distance between the hashes of two vectors ceases to provide any information about the true distance between the vectors. We will find this property useful in developing our privacy-preserving system. Changing the value of the precision parameter Δ allows us to adjust the distance threshold until which the Hamming distance is informative. Increasing the number of bits M leads to a reduction of the variance of the Hamming distance. A converse

property of the embeddings is that for all $\mathbf{x}'$ except those that lie within a small radius of any $\mathbf{x}$, $d_H(\mathbf{q}(\mathbf{x}), \mathbf{q}(\mathbf{x}'))$ provides little information about how close $\mathbf{x}'$ is to $\mathbf{x}$. It can be shown that the embedding provides information theoretic security beyond this radius, if the embedding parameters $\mathbf{A}$ and $\mathbf{w}$ are unknown to the potential eavesdropper. Any algorithm attempting to recover a signal $\mathbf{x}$ from its embedding $\mathbf{q}(\mathbf{x})$ or to infer anything about the relationship between two signals sufficiently far apart using only their embeddings will fail to do so. Furthermore, even in the case where $\mathbf{A}$ and $\mathbf{w}$ are known, it seems computationally intractable to derive $\mathbf{x}$ from $\mathbf{q}(\mathbf{x})$ unless one can guess a starting point very close to $\mathbf{x}$. In effect, it is infeasible to invert the SBE without strong *a priori* assumptions about $\mathbf{x}$.

5. SECURE IMPORTANT PASSAGE RETRIEVAL

Our approach for a privacy-preserving important passage retrieval system closely follows the formulation presented in Section 3, and it is illustrated in Figure 4. However, this is a very important difference in terms of who performs each of the steps. Typically there is only one party involved, Alice, who owns the original documents, performs key phrase extraction, combines them with the bag-of-words model in a compact matrix representation, computes the support sets for each documents and finally uses to retrieve the important passages. In our scenario, Alice does not know how to extract the important passages from them or does not possess the computational power to do so. Therefore, she must outsource the retrieval process to a third-party, Bob, who has these capabilities. However, Alice must first obfuscate the information contained in the compact matrix representation. If Bob receives this information as is, he could use the term frequencies to infer on the contents of the original documents and gain access to private or classified information Alice does not wish to disclosure to anyone. Alice computes binary hashes of her compact matrix representation using the method described in Section 4, keeping the randomization parameters $\mathbf{A}$ and $\mathbf{w}$ to herself. She sends these hashes to Bob, who computes the support sets and extracts the important passages. Because Bob receives binary hashes instead of the original matrix representation, he must use the normalized Hamming distance instead of the cosine distance in this step, since it is the metric the SBE hashes best relate to. Finally, we returns the hashes corresponding to the important passages to Alice, who then uses them to get the information she desires.

6. EXPERIMENTS

In this section we illustrate the performance of our privacy-preserving approach to Important Passage Retrieval and how it compares to its non-private counterpart. We start by presenting the datasets we used in our experiments, then we describe the experimental setup and finally we present some results.

6.1 Datasets

In order to evaluate our approach, we performed experiments on the English version of the Concisus dataset [26] and the Portuguese Broadcast News (PT BN) dataset [23]. The Concisus dataset is composed by seventy eight event reports and respective summaries, distributed across three different types of events: aviation accidents, earthquakes, and train accidents. This corpus also contains comparable data in Spanish. However, since our Automatic Key Phrase Extraction (AKE) system uses some language-dependent features, we opted for not using in this part of the dataset in previous work [24] nor in this one.

The PT BN dataset consists of automatic transcriptions of eighteen broadcast news stories in European Portuguese, which are part of a news program. News stories cover several generic topics like society, politics and sports, among others. For each news story, there is a human-produced abstract, used as reference.

Metric	ROUGE-1
Cosine distance	0.575
Euclidean distance	0.507

Table 1: KP-Centrality results with 40 key phrases using the Concisus dataset.

Metric	ROUGE-1
Cosine distance	0.612
Euclidean distance	0.590

Table 2: KP-Centrality results with 40 key phrases using the Portuguese Broadcast News dataset.

6.2 Setup

We extracted key phrases from both datasets using the MAUI toolkit [16] expanded with shallow semantic features, such as number of named entities, part-of-speech tags and four n-gram domain model probabilities. This expanded feature set leads to improvements regarding the quality of the key phrases. Regarding the Concisus dataset, we extracted yet additional features, such as the detection of rhetorical devices, which further improved the key phrase extraction process [13]. As for the PT BN dataset, we only used the shallow semantic features as the remaining features were not available [14].

6.3 Results

We present some baseline experiments in order to obtain reference values for our approach. We generated three passages summaries for Concisus Dataset, which are commonly found in online news web sites like Google News. In the experiments using the PT BN dataset, the summary size was determined by the size of the reference human summaries, which consisted in about 10% of the input news story. For both experiments, we used the Cosine and the Euclidean distance as evaluation metrics, since the first is the usual metric for computing textual similarity, but the second is the one that relates to the Secured Binary Embeddings technique. All results are presented in terms of ROUGE [10], in particular ROUGE-1, which is the most widely used evaluation measure for this scenario. The results we obtained for the Concisus and the PT BN datasets are presented in Tables 1 and 2, respectively.

We considered forty key phrases in our experiments since it is the usual choice when news articles are considered [13]. As expected, we notice some slight degradation when the Euclidean distance is considered, but we still achieve better results than other state-of-the-art methods such as default centrality [22] and LexRank [5]. Reported results in the

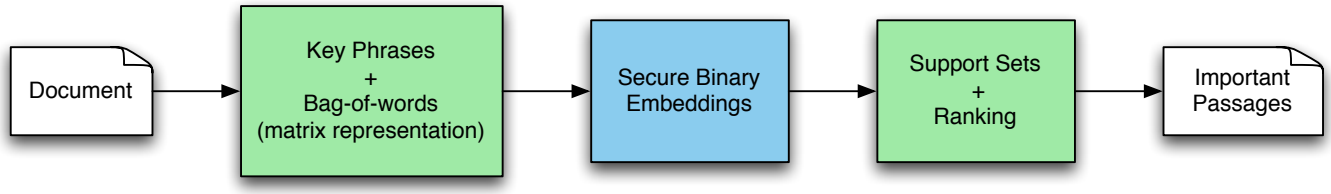

Figure 4: Flowchart of the Secure Important Passage Retrieval method.

leakage	∼ 5%	∼ 25%	∼ 50%	∼ 75%	∼ 95%
bpc=4	0.365	0.437	0.465	0.486	0.495
bpc=8	0.424	0.384	0.436	0.452	0.500
bpc=16	0.384	0.416	0.450	0.463	0.517

Table 3: KP-Centrality using SBE and the Concisus dataset, in terms of ROUGE-1.

leakage	∼ 5%	∼ 25%	∼ 50%	∼ 75%	∼ 95%
bpc=4	0.314	0.340	0.470	0.478	0.562
bpc=8	0.327	0.324	0.486	0.507	0.527
bpc=16	0.338	0.336	0.520	0.473	0.550

Table 4: KP-Centrality using SBE and the Portuguese Broadcast News dataset, in terms of ROUGE-1.

literature include ROUGE-1 = 0.443 and 0.531 using default Centrality and ROUGE-1 = 0.428 and 0.471 using LexRank for the Concisus and PT BN datasets, respectively [24]. This means that the forced change of metric due to the intrinsic properties of SBE does not affect the validity of our approach in any way.

For our privacy-preserving approach we performed experiments using different values for the SBE parameters. The results we obtained in terms of ROUGE for the Concisus and the PT BN datasets are presented in Tables 3 and 4, respectively. Leakage refers to the fraction of SBE hashes for which the normalized Hamming distance d_H is proportional to the Euclidean distance d_E between the original data vectors. The amount of leakage is exclusively controlled by Δ. Bits per coefficient (*bpc*) is the ratio between the number of measurements M and the dimensionality of the original data vectors L, i.e., $bpc = M/L$. As expected, increasing the amount of leakage (i.e. increasing Δ) leads to improving the retrieval results. Surprisingly, changing the values of *bpc* does not lead to improved performance. The reason for this results might be due to the KP-Centrality method using support sets that consider multiple partial representations of the documents. Nevertheless, the most significant results is that for 95% leakage there is an almost negligible loss of performance. This scenario, however, does not violate our privacy requisites in any way, since although most of the distances between hashes are known, there is no way to use this information to learn anything about the original underlying information.

7. CONCLUSIONS AND FUTURE WORK

In this work, we introduced a privacy-preserving technique for performing Important Passage Retrieval that performs similarly to their non-private counterpart. Our Secure Binary Embeddings based approach provides secure document representations that allows for sensitive information to be processed by third parties without any risk of sensitive information disclosure. Although there was some slight degradation of results regarding the baseline, our approach still outperforms other state-of-the-art methods like default Centrality and LexRank, but with important advantage that no private or classified information is disclosed to third parties.

For future work we intend to use the secure representation based on Secure Binary Embeddings in multi-document important passage retrieval. Another additional research line that we would like to purse is to apply this privacy preserving technique in other Information Retrieval tasks, such as classified military and medical records retrieval.

8. ACKNOWLEDGMENTS

We would like to thank FCT for supporting this research through PPEst-OE/EEI/LA0021/2013, the Carnegie Mellon Portugal Program, PTDC/EIA-CCO/122542/2010, and grants SFRH/BD/33769/2009 and SFRH/BD/71349/2010. We would like to thank NSF for supporting this research through grant 1017256.

9. REFERENCES

[1] M. Bellare, V. T. Hoang, S. Keelveedhi, and P. Rogaway. Efficient garbling from a fixed-key blockcipher. In *IEEE Symposium on SP*, pages 478–492. IEEE, 2013.

[2] P. Boufounos. Universal rate-efficient scalar quantization. *IEEE Transactions on Information Theory*, 58(3):1861–1872, 2012.

[3] P. Boufounos and S. Rane. Secure binary embeddings for privacy preserving nearest neighbors. In *WIFS 2011*, pages 1–6. IEEE, 2011.

[4] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR 1998: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM, 1998.

[5] G. Erkan and D. R. Radev. LexRank: Graph-based Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[6] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

[7] W. Jiang and B. Samanthula. N-gram based secure similar document detection. In Y. Li, editor, *Data and Applications Security and Privacy XXV*, volume 6818

of *Lecture Notes in Computer Science*, pages 239–246. Springer Berlin Heidelberg, 2011.

[8] W. Jiang, L. Si, and J. Li. Protecting source privacy in federated search. In *SIGIR 2007: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 761–762, New York, NY, USA, 2007. ACM.

[9] T. K. Landauer and S. T. Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psych. Review*, 104(2):211–240, 1997.

[10] C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In M.-F. Moens and S. Szpakowicz, editors, *Text Summ. Branches Out: Proc. of the ACL-04 Workshop*, pages 74–81. Association for Computational Linguistics, 2004.

[11] M. Litvak and M. Last. Graph-Based Keyword Extraction for Single-Document Summarization. In *Coling 2008: MMIES*, pages 17–24. Coling 2008 Org. Committee, 2008.

[12] W. Lu, A. Varna, and M. Wu. Confidentiality-preserving image search: A comparative study between homomorphic encryption and distance-preserving randomization. *Access, IEEE*, 2:125–141, 2014.

[13] L. Marujo, A. Gershman, J. Carbonell, R. Frederking, and J. P. Neto. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In *Proceedings of the Eighth International Language Resources and Evaluation (LREC 2012)*, 2012.

[14] L. Marujo, M. Viveiros, and J. P. Neto. Keyphrase Cloud Generation of Broadcast News. In *Interspeech 2011*. ISCA, September 2011.

[15] S. R. Maskey and J. Hirschberg. Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization. In *Proceedings of the 9^{th} EUROSPEECH - INTERSPEECH 2005*, 2005.

[16] O. Medelyan, V. Perrone, and I. H. Witten. Subject metadata support powered by Maui. In *Proceedings of the JCDL '10*, page 407, New York, USA, 2010.

[17] M. Murugesan, W. Jiang, C. Clifton, L. Si, and J. Vaidya. Efficient privacy-preserving similar document detection. *The VLDB Journal*, 19(4):457–475, 2010.

[18] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *EUROCRYPT'99*, pages 223–238, 1999.

[19] H. Pang, X. Xiao, and J. Shen. Obfuscating the topical intention in enterprise text search. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1168–1179, April 2012.

[20] G. Penn and X. Zhu. A Critical Reassessment of Evaluation Baselines for Speech Summarization. In *Proc. of ACL-08: HLT*, pages 470–478. ACL, 2008.

[21] J. Portelo, B. Raj, P. Boufounos, I. Trancoso, and A. Alberto. Speaker verification using secure binary embeddings. In *EUSIPO*, 2013.

[22] R. Ribeiro and D. M. de Matos. Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. *Journal of Artificial Intelligence Research*, 42:275–308, 2011.

[23] R. Ribeiro and D. M. de Matos. *Multi-source, Multilingual Information Extraction and Summarization*, chapter Improving Speech-to-Text Summarization by Using Additional Information Sources. Theory and Applications of NLP. Springer, 2013.

[24] R. Ribeiro, L. Marujo, D. Martins de Matos, J. P. Neto, A. Gershman, and J. Carbonell. Self reinforcement for important passage retrieval. In *SIGIR 2013: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 845–848. ACM, 2013.

[25] K. Riedhammer, B. Favre, and D. Hakkani-Tür. Long story short – Global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52:801–815, 2010.

[26] H. Saggion and S. Szasz. The concisus corpus of event summaries. In *Proceedings of the Eighth International Language Resources and Evaluation (LREC 2012)*, 2012.

[27] R. Sipos, A. Swaminathan, P. Shivaswamy, and T. Joachims. Temporal corpus summarization using submodular word coverage. In *Proc. of CIKM*, pages 754–763, New York, NY, USA, 2012. ACM.

[28] R. I. Tucker and K. Spärck Jones. Between shallow and deep: an experiment in automatic summarising. Technical Report 632, University of Cambridge, 2005.

[29] V. R. Uzêda, T. A. S. Pardo, and M. das Graças Volpe Nunes. A comprehensive comparative evaluation of RST-based summarization methods. *ACM Trans. on Speech and Language Processing*, 6(4):1–20, 2010.

[30] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova. Beyond SumBasic: Task-focused summarization and lexical expansion. *Information Processing and Management*, 43:1606–1618, 2007.

[31] X. Wan, J. Yang, and J. Xiao. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th ACL*, pages 552–559, 2007.

[32] A. C.-C. Yao. Protocols for secure computations. In *Foundations of Computer Science (FOCS)*, volume 82, pages 160–164, 1982.

[33] K. Zechner and A. Waibel. Minimizing Word Error Rate in Textual Summaries of Spoken Language. In *Proceedings of the North American Chapter of the ACL (NAACL)*, pages 186–193, 2000.

[34] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *SIGIR 2002: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–120. ACM, 2002.

[35] J. J. Zhang, R. H. Y. Chan, and P. Fung. Extractive Speech Summarization Using Shallow Rhetorical Structure Modeling. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 18(6):1147–1157, 2010.