Providing Privacy in Domain Specific Search with SSM and Cosine Similarity

[1]Ms.Suvarna A Veer, [2]Rajani S. Sajjan
[1]Student, [2]Professor
Computer Science & Engineering Department,
VVPIET, Solapur, India

Abstract: Personalized web search (PWS) improves the level of various search services on the Internet. But because of lack of user's private search information during search it is difficult for the wide proliferation of PWS. We study privacy protection in PWS applications which helps to structure user preferences as hierarchical user profiles. We propose a PWS framework that helps to generalize profiles by queries while respecting user specified privacy requirements. We are aiming at striking a balance between two main factors that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. We present two algorithms first is SSM and cosine similarity algorithm for runtime generalization and second is ranking based on mean value of similar links.

IndexTerms **- Data security, public server, SSM, PWS**

I. INTRODUCTION

The web search engine has long turn into the most vital source for individuals searching for helpful data on the web. However, users may get disappointed when web search return unwanted results that don't meet their requirements. Such unwanted data extracted by search engine is generally because of huge amount of variety of users' contexts and backgrounds, and additionally the ambiguity of writings. Personalized Web Search (PWS) is a general class of pursuit procedures going for giving better indexed lists, which are custom-made for individual client needs[5]. As the cost, user's data must be classified to make sense of the client expectation behind the issued inquiry.

The PWS can be classified as click-log-based systems and profile-based systems. The click log-based strategies basically go through clicked pages in the client's search history[4]. Despite the fact that this system has been exhibited to perform reliably and extensively well, it can just chip away at new domain inquiries from the same client, which makes it difficult to keeping of its appropriateness. On the other side, profile-based system generates the client interest models created from client profiling strategies. Although both types of PWS techniques has some advantages and disadvantages for, the profile-based PWS has proved more effective in improving the quality of web search recently, with increasing usage of personal and behavior information to profile its users, which is usually gathered from query history browsing history, click-through data, bookmarks, user documents and so forth. Unfortunately, such certainly collected personal data can easily reveal a matter of interest of user's private life [7].

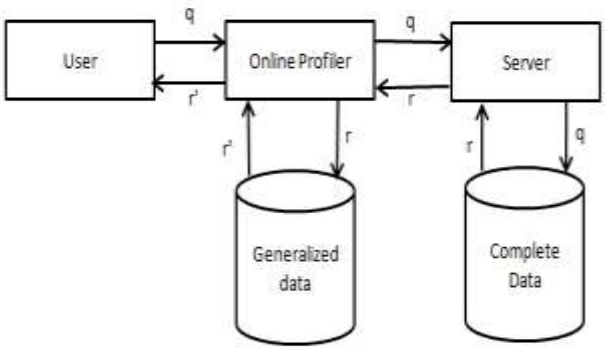

Fig 1.1 Generalized user profiler

The Web personalization process is classified into four deferent phases: Collection of Web data - Implicit data includes past activities as recorded in Web server logs via cookies or session tracking modules. Explicit data generally gathered from registration forms and rating questionnaires. Additional data such as demographic and application data (for example, e-commerce transactions) can also be used [1][4]. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages. Pre-processing of Web data – Gathered data is further pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step.

Pre-processing may include cleaning data of inconsistencies, filtering out irrelevant information according to the goal of analysis (example: automatically generated requests to embedded graphics will be recorded in web server logs, even though they add little information about user interests), and completing the missing links (due to caching) in incomplete click through paths [4]. Analysis of Web data - Also called Web Usage Mining, this step applies data mining techniques to discover interesting usage

patterns and statistical correlations between web pages and user groups. This step generates user profile, and work offline, so that it does not add any extra workload on the web server. Final Recommendation Phase -This is the last phase; this step uses the results of the previous analysis step to provide recommendations to the user. The recommendation process involves generating dynamic web content and adding hyperlinks to the last web page requested by the user [3][6].

II. LITERATURE SURVEY

[1] In this paper author proposed an efficient information retrieval system in order to overcome the drawbacks of the ranking algorithms and improve the efficiency of web searching respecting to the precision measures. Current search engines do not rank the searched documents for a certain query automatically; they just retrieve related documents to that query issued by the user.

[2] Author proposed a framework called Personalized Mobile Search Engine (PMSE) which extracts and learns user's search and location preferences based on the user's clickthrough. The GPS trajectories are used to adapt the user mobility. GPS locations help to improve effectiveness during retrieval, especially for location queries. Two privacy parameters, minDistance and expRatio are proposed. The privacy parameters provide implicit control of privacy exposure while maintaining good ranking quality.

[3] In this paper author implemented system a client-side privacy protection framework. System is potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The framework allowed users to specify customized privacy requirements via the hierarchical profiles. In addition, Online generalization on user profiles to protect the personal privacy without compromising the search quality. GreedyDP and GreedyIL algorithm for the online generalization. Our experimental results revealed that system could achieve quality search results while preserving user's customized privacy requirements. The results also confirmed the effectiveness and efficiency of our solution.

[5] have proposed a strategy for recommending the preferences directly for personalized web searches to increase the appropriate- ness of the results. A unique approach of conditional preference networks is employed for recommending preferences more explicitly. The approach directly aims at improving the overall accuracy of the personalized web search. The user preferences are first obtained which act as the intrinsic driving force for recommending web pages to the users.

Ramya and Gowthami [9] have proposed a per- sonalized web search mechanism by implementing the meta search approach that relies on the existing meta search engines like Google, Bing, Yahoo, etc. The client receives the request from the users' and submits to the server and displays the results based on his/her profile details and favorite search history. The server manages the tasks and forwards the request to search engines. The user details are stored in the user profile that pre- serves the privacy. That makes client-server model to communicate in a faster way and provides more efficiency in results based on the user query.

Abu-Dalbouh [10] has conducted a detailed study on incorporating end- user privacy in personalized web search systems. A study focuses on facilitating an end to end privacy in human computer systems for achieving better retrieval privacy in Web Search Systems. The main inference from the study is that privacy is a criterion which must accompany the Web Searches, as a Web Search without privacy is incomplete and risky.

Table 1.1 shows the comparison between similarity algorithms with similarity and time parameters by evaluating results on string 1 and string 2. Here Levenshtein distance algorithm takes considerable amount of time with minimum similarity. Cosine similarity gives good result with higher similarity rate with minimum time.

String 1: Android mobile development company

String 2: What is android?

ALGORITHM	SIMILARITY (%)	TIME(MS)
Levenshtein distance [11]	14.71%	2500
Q-grams [12]	14.55%	2320
Dice coefficient [13]	28.57%	2360
Cosine similarity [14]	28.87%	1200

Table 1.1 String similarity algorithms comparisons

III. PROBLEM STATEMENT

In the proposed system, we are going to implement the process by using which the system can become capable of capturing and extracting a series of queries by applying string similarity match algorithm to minimize the computational time and to achieve more accuracy in search results.

IV. PROPOSED SYSTEM ARCHITECTURE

Web search engines (e.g. Google, Yahoo, Microsoft Live Search, Bing, etc.) are mostly used to search certain information from a large amount of data in a very few amounts of time [7]. These web search engines also pose a privacy threat to the users. These useful tools profile their users on the basis of previous searches submitted by them. To address this privacy threat, current solution proposes new mechanisms that introduce a low cost in terms of communication. In the proposed system, we are going to implement the String Similarity Match Algorithm (SSM Algorithm) for increasing the better search quality results. Personalized web search is best way to increase the accuracy of web search.

In the proposed system, we propose a new protocol specially designed to provide privacy to the user's in front of web search profiling. In this proposed system, we propose and try to oppose adversaries with broader background knowledge, such as richer relationship among topics. We have generalized the user profile results by using the background knowledge which is going to store in history. Through this we are able to hide the user search results. In the existing system, Greedy IL and Greedy DP algorithm are used which takes large computational time.

In the proposed system, Cosine Similarity and String Similarity Match Algorithm is used. For Client post query q to the server, server retrieves the query results PT to the client. Results have been extracted by using string similarity match algorithm. One possible definition of the string similarity match algorithm is the following:

1. Given a pattern string, P = p1p2…..pm
2. Text string , T = t1t2……..tn
3. Find a substring Tj',j = tj'…….tj
   in T, which of all substrings of T, has the smallest edit distance to the pattern P.

4.1 Algorithm for Proposed System:

Step 1: Detection & removal of unwanted symbols
Step 2: Calculate similarity values for user given words and words in database.
Step 3: In that similarity computation, extract the similar features in the dataset.
Step 4: Then calculate the ASCII difference for user given word and words in database.
Step 5: Then again calculate the similarity values.
Step 6: Finally retrieve the most relevant documents based on the similar matching values.

Psuedocode:

```
public float DistanceCalculation(String source, String target)
{
final int sl = source.length();
final int tl = target.length();

if (sl == 0 || tl == 0) {
if (sl == tl) {
return 1;
}
else {
return 0;
}   }

int cost = 0;
if (sl < n || tl < n) {
for (int i=0,ni=Math.min(sl,tl);i<ni;i++) {
if (source.charAt(i) == target.charAt(i)) {
cost++;
}
}
return (float) cost/Math.max(sl, tl);
}
```

Above pseudocode gives the distance between two strings namely source string and target string. Algorithm performs character wise string similarity. Algorithm checks character at same index location in both strings, if character matches then it increases cost by 1 if not it goes to next index position till end of word. Finally, distance is calculated by dividing cost with maximum string length.
The proposed framework will dynamically generate a user profile for a user's query prioritizing the user's privacy.

Search Query:
Here client execute the Query (q) to the server, server retrieves the data and retrieved data is further generalized and similarity between dataset keyword and data extracted from server is calculated. Weights are assigned to each links as per similarity value.

SSM and Cosine Similarity with TFIDF:

Module 1: Processing of multiple links :

1. **Coreferent Detection:**

- Removal of Duplicates.
- Tokenization.
- Filtering

Module 2: Cosine Similarity Based Approach:

Cosine Similarity measures the similarity between two sentences in terms of the value within the range of [0, 1]. Cosine similarity is based on basic fundamentals o term based extraction. Term frequency identifies the links containing the relevant terms TF (term, document) = Frequency of term / No of links

$$tf_i = \frac{n_i}{\sum_k n_k} \quad (1)$$

IDF (inverse document frequency) calculates whether the term is rare or common in all documents. IDF (term, document) is calculated by dividing total number of documents by the number of documents containing provided term and taking log of that.(Here documents are nothing but the links)

IDF (term, no of answers) = log (Total No of links / No of links containing term)

$$idf_i = \log \frac{D}{|\{d:t \in d\}|} \quad (2)$$

TF-IDF is the multiple of the value of TF and IDF for a provided term. The increasing value of TF-IDF is directly proportional to the number of occurrences of term within a link and with rarity of the term across the corps

TFIDF=TF*IDF (3)

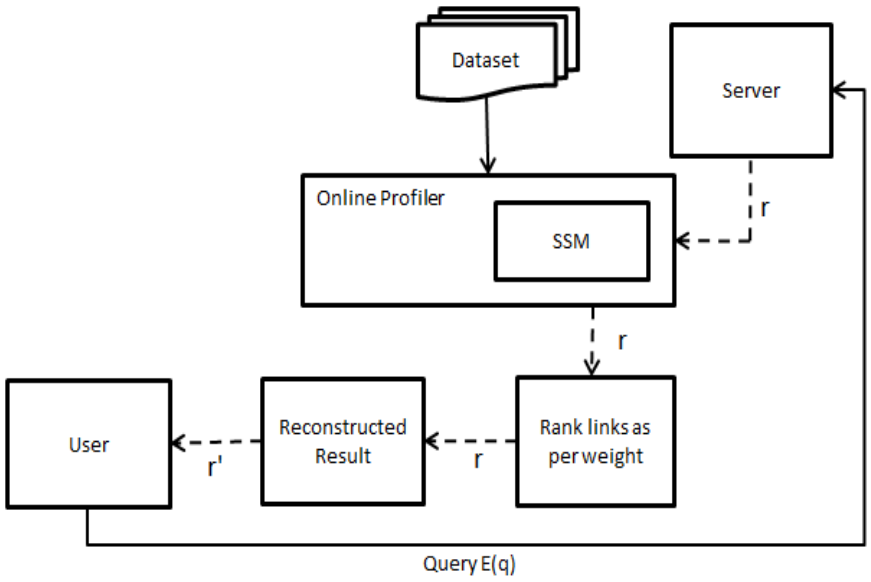

Fig. 4.1 Proposed System Architecture

Module 3: Ranking Algorithm Implementation:

Ranking Algorithm calculates the mean of whole links, creates a mean value and it takes nearest value from the mean value and generates the output based on score.

Steps-

Arrange the Value in descending order.

Calculate Mean Value for whole Answers.

$Mean\ Value$= Total no. of links /Total Size

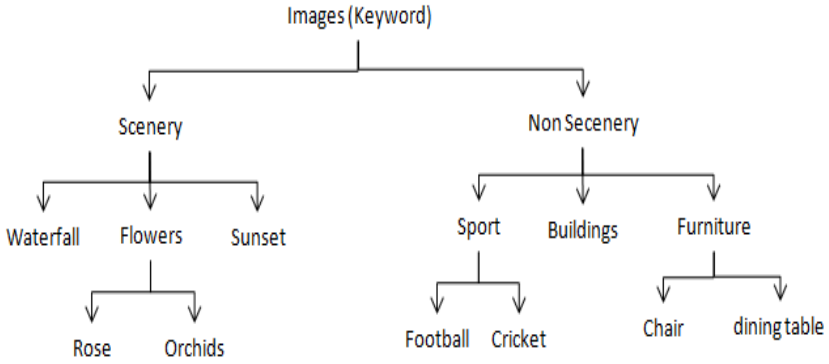

Fig 4.2 User Profile

4.2 DATA FLOW DIAGRAM (DFD):

This Data Flow involves following steps -

1. **Data Pre-processing Phase-**
   Retrieve data
   Eliminate stop words.
2. **Feature extraction Phase-**
   Extract relevant feature i.e weighted feature.
3. **Similarity Based Approach Phase-**
   Rank similar multiple links

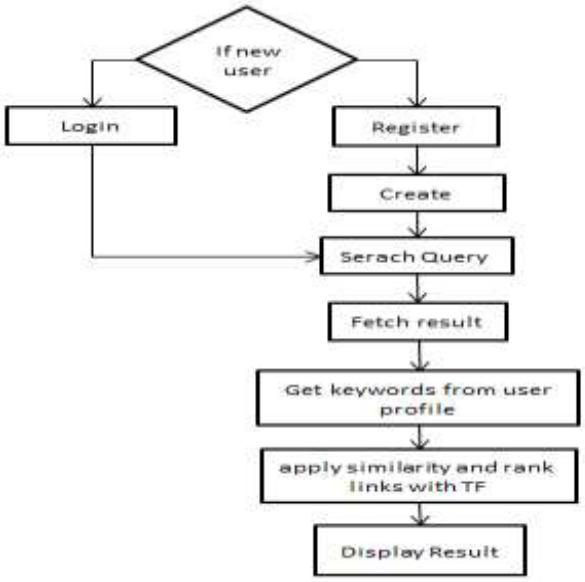

Fig. 4.3 System flow diagram

4.3 ADVANTAGES

1. It provides better search results.
2. It has less computational time as compared to existing system.
3. It is simple
4. It is very efficient to evaluate.

CONCLUSION

The data can be retrieved by using the background knowledge for generalization. An important feature of transaction data is the extreme sparsity, which makes any single technique not sufficient in anonymizing such data. Among recent works, some suffer from high information loss, some result in data hard to interpret, and some suffer from performance drawbacks. From some previous studies, it can be seen that most of the users are willing to compromise privacy if the personalization by supplying user profile to the search engine provides better search quality. In the proposed system, we propose generalization to minimize information loss. We propose new techniques to address the efficiency and scalability challenges.

Proposed system provides better quality results and provides more efficiency. Our string similarity match algorithm provides better accuracy.

Acknowledgment

The authors would like to thank the researchers as well as publishers for making their resources available and also thanks to the Director of VVPIET, Principal, HOD of Computer Department, PG Coordinator, Project Guide and Co-Guide and faculties for constant encouragement.

REFERENCES

[1] S. Manek fmanek3@gmail.com Aishwarya J. Reddy Vaibhavu panchal Vijaya Pinjarkar Department of Information Technology K.J.Somaiya Institute of Engineering and Information Technology, Sion. Mumbai, Maharashtra, India ykhirodkar@somaiya.edu "Hybrid Crawling for Time-Based Personalized Web Search Ranking Foram" 978-1-5090-5686-6/17/$31.00 ©2017 IEEE

[2] Pratibha Rathod pratima.rathod15@gmail.com Smita Desmukh Information Technology, deshmukhsmita17@yahoo.com "A Personalized Mobile Search Engine based on User Preference" IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017)

[3] Mrs. Sharvari V. Malthankar , Prof. Shilpa Kolte PG Student,"Client side Privacy Protection Using Personalized Web Search" Elsevier Science direct 7th International Conference on Communication, Computing and Virtualization 2016

[4] LidanShou, He Bai, Ke Chen, and Gang Chen, "Supporting Privacy Protection in PersonalizedWeb Search", IEEE Transactions on Knowledge and Data engineering,vol.26,no.2,february 2014

[5] Sachin S. Kale1, Dattatray N. Udmale1, Anjali B. Navale1, Prerana S. Wagh1, Prof. Rahinj P.L2 B.E Supporting Privacy Protection in Personalized Web Search, India International Journal of Innovative Research in Computer and Communication Engineering 2017

[6] Brahmaji Katragadda, 2Sk.Meera "Supporting Privacy Protection in Personalized Web Search" Journal of Science and Technology (JST) Volume 2, Issue 7, July 2017

[7] Anoj Kumar anoj.kr@hotmail.com Mohd. Ashraf ashraf.saifee@gmail.com Personalized Web Search Engine using Dynamic User Profile and Clustering Techniques" 978- 9-3 805-4416-8/15/$31. 00 c2 01 5 IEEE

[8] K R Remesh Babua,Philip Samuelb, "Concept Networks for Personalized Web Search Using Genetic Algorithm" International Conference on Information and Communication Technologies 2016

[9] Kamlesh Makvana, Pinal Shah, Parth Shah, kamleshmakvana.it@charusat.ac.in, parthshah.ce@charusat.ac.in pinalshah.it@charusat.ac.in "A Novel Approach to Personalize Web Search through User Profiling and Query Reformulation" 978-1-4799-4674-7/14/$31.00©2014 IEEE

[10] V.Ramya, S.Gowthami "ENHANCE PRIVACY SEARCH IN WEB SEARCH ENGINE USING GREEDY ALGORITHM" International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 3, Issue 8, November 2014

[11] Zhan Su, Byung-Ryul Ahn, Ki-yol Eom, Min-koo Kang, Jin-Pyung Kim, Moon-Kyun Kim Department of Artificial Intelligence, University of Sungkyunkwan Cheoncheon dong, Jangan-gu, Suwon, Korea Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm The 3rd Intetnational Conference on Innovative Computing Information and Control (ICICIC'08) 978-0-7695-3161-8/08 $25.00 © 2008 IEEE

[12] LEENA SALMELA, JORMA TARHIO and JARI KYT¨ OJOKI Helsinki University of Technology ACM Journal 2008

Feddy Setio Pribadi1,2, Teguh Bharata Adji1, Adhistya Erna Permanasari1 1Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Indonesia Automated Short Answer Scoring Using Weighted Cosine Coefficient, 2016 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.

[13] Alfirna Rizqi Lahitani1), Adhistya Erna Permanasari2), Noor Akhmad Setiawan3) (123Department of Electrical Engineering and Information Technology, Faculty of Engineering Universitas Gadjah Mada Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment ACM 2011

[14] ZHENGHUA XU1 , OANA TIFREA-MARCIUSKA2 , THOMAS LUKASIEWICZ1 ,MARIA VANINA MARTINEZ3 , GERARDO I. SIMARI3 , and CHENG CHEN "Lightweight Tag-Aware Personalized Recommendation on the Social Web Using Ontological Similarity" 4 2169-3536 (c) 2018 IEEE

[15] Yayuan Tang 1;2 , Hao Wang 3 , Kehua Guo 2;4 , Yizhe Xiao 2 , Tao Chi "Relevant Feedback Based Accurate and IntelligentRetrieval on Capturing User Intention for Personalized Websites" 2169-3536 (c) 2018 IEEE

[16] Mohammad Mustaneer Rahman, and Nor Aniza Abdullah, "A Personalised Group-Based Recommendation Approach for Web Search in E-Learning" IEEE 2169-3536 (c) 2018 IEEE.

[17] Puxuan Yu Wuhan University Wuhan, China pxyuwhu@gmail.com Wasi Uddin Ahmad wasiahmad@ucla.edu "Hide-n-Seek: An Intent-aware Privacy Protection Plugin for Personalized Web Search" 18, July 8-12, 2018, Ann Arbor, MI, USA Italy. 2016 ACM. ISBN 978-1-4503-4069-4/16/07

[18] Gerard Deepak, B. N. Shwetha, C. N. Pushpa, J. Thriveni & K. R. Venugopal "A hybridized semantic trust-based framework for personalized web page recommendation" International Journal of Computers and Applications ISSN: 1206-212X (Print) 1925-7074

[19] Najneen Tamboli, Sathish Kumar "Revi ew on Pri vacy Preservation in Person al i z ed Web Search" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 11, November 2015

[20] Avi Arampatzis, George Drosatos and Pavlos S . Efraimidis," A Versatile Tool for Privacy -Enhanced Web Searc" Xant hi 67 100, Greece Springer- Verlag Berlin Heidelberg 2013

[21] X.Smilien Rophie, smilienrophie@gmail.com Dr. A. Anitha, dr.aanitha@yahoo.com "A Versatile Tool for Privacy -Enhanced Web Search" International Journal of Technology and Engineering System (IJTES) Vol 8. No.1 – Jan-March 2016 Pp. 44-49

[22] B. SekharBabu, P. Lakshmi Prasanna, D. Rajeswara Rao, J. LakshmiAnusha, A. Pratyusha and A. Ravi Chand, "PROFILE BASED PERSONALIZED WEB SEARCH USING GREEDY ALGORITHMS" MAY 2016 ISSN 1819-6608 ARPN Journal of Engineering and Applied Sciences

[23] M. Spertta and S. Gach, "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.

[24] Kan Yang, and Xiaohua Jia, " Expressive, Efficient, and Revocable Data Access Control for Multi-Authority Cloud Storage", *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS,*Vol.25, No.7, pp. 1735-1744, July 2014

[25] Ming Li, Member , IEEE, Shucheng Yu, Member , IEEE , Yao Zheng, Student Member , IEEE , Kui Ren, Senior Member , IEEE , and Wenjing Lou, Senior Member , IEEE "Scalable and Secure Sharing of Personal Health Records in Cloud Computing Using Attribute-Based Encryption*" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 1, JANUARY 2013*

[26] Syam Kumar P, Subramanian R Department of Computer Science, School of Engineering & Technology Pondicherry University "An Efficient and Secure Protocol for Ensuring Data Storage Security in Cloud Computing*" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 1, November 2011*

[27]Cong Wang, Qian Wang, and Kui Ren Department of ECE Illinois Institute of Technology "Ensuring Data Storage Security in Cloud Computing" *Email: cwang, qwang, Wenjing Lou Department of ECE Worcester Polytechnic Institute JULY 2012* C. Rolim, F. Koch, C. Westphall, J. Werner, A. Fracalossi, and G. Sal- vador, "A Cloud Computing Solution for Patient's Data Collection in Health Care Institutions," in Proc. ETELEMED, 2010.