Comparative Studies of Information Retrieval Approaches in User-Centered Health Information System

Ibrahim Umar Kontagora[1,2] and Isredza Rahmi A. Hamid[1](✉)

[1] Faculty of Computer Science and Information Technology, Information Security Interest Group (ISIG), Universiti Tun Hussein Onn, Parit Raja, Malaysia
ibrosoftuk@yahoo.com, rahmi@uthm.edu.my
[2] Department of Computer Science, Niger State Polytechnic, Zungeru, Niger State, Nigeria

Abstract. In this paper, a comparative studies of different methods deployed in addressing problems of user-centered health information retrieval systems were investigated. The reason for the comparative studies is to identify the approach that best addressed the readability and vocabulary mismatched issues encountered by laymen patients and their relatives in exploring information extracted from medical discharge documents and clinical reports online. We discussed and presented the performance of information retrieval systems in previous research works. We concentrated on classifying and comparing the three approaches used in health information retrieval which are Vector Space Model (VSM), Language Based Approach Model (LM) and Context Based Approach (CBA). The usefulness of incorporating controlled vocabularies such as Metamap, UMLS, external, MeSH, etc. was extensively discussed. The result shows that the Language Based Approach systems achieved better results as compared to the Vector Space Model Approach and Context Based Approach Systems. The Language Based Approach Systems managed to acquire 0.4146, 0.7560 and 0.7445 for Mean Average Precision, Precision @ 10 and Normalized Discontinued Cumulative Gains @ 10 respectively. Hence, we conclude based on the outcome of the comparative studies and our experimental results that the language modeling models is best suited to be deployed in addressing the problems of returning relevant information by user centered health information retrievals to users.

Keywords: Language models · Vector space models concept-based approach External medical resource · Query expansion

1 Introduction

Health information retrieval has become attractive today as a result of the massive rise in medical allied information online [1]. There is no system at present that doubts the information needs of the user context and returns appropriate documents to the best of their knowledge [2]. With the spreading consciousness, probing online health related network mediums and other sources has become a frequent habit [3]. A study

R. Ghazali et al. (eds.), *Recent Advances on Soft Computing and Data Mining*,
Advances in Intelligent Systems and Computing 700,
https://doi.org/10.1007/978-3-319-72550-5_17

conducted by Pew Research Center [4] revealed that a huge percentage of the population in the United States' search engine operators seek for information on a particular illness online.

The previous campaigns addressed user different information needs. Moreover, it only targeted on a specific group of consumers with skilled health awareness (e.g., health researchers and clinicians) [5]. The ShARe/CLEF eHealth Task 3 launched an information retrieval campaign which focused on the information needs of common people and the queries they created to express their needs. However, the outcome of the (2013) Information retrieval tasks failed to tackle patients' questions which they may come up while reading the clinical report [1]. This same result has showed significant progress over the 2012 results for both the reference line and team submissions using the new query set [3].

The main objective of this study is to carry out comparative studies on the various information retrieval approaches deployed to address the problems of user centered health information retrievals. Hence, we are going to propose a new information retrieval approach for health information system. The remainder of this paper is organized as follows. Section 2 describes related work regarding Information retrieval approaches. Section 3 discusses the comparison of the approaches deployed in addressing information retrieval problems, Sect. 4 contains the result and discussions and Sect. 5 concludes the work and direction for future work.

2 Related Work

Medical Information Retrieval is widely measured as a health related assignment usually executed by a huge range of medical workers and laypeople (patients and their relatives) [6]. Nearly 80% of United States search engine operators' searched for health related facts about a particular illness through online [5]. Various prospective information seekers with different medical understanding, suggests the increased amount of information required. Subsequently, the medical information retrieval systems design requirement must meet up with the health information needs of different categories of users [7].

Patients are always eager of knowing the exact content of their discharge summaries written by medical expert [6]. However, the medical text is highly professional for a layman to follow [1]. Therefore, the medical information retrieval becomes highly accepted as helpful way in answering the patient's questions [7]. Number of evaluation campaigns focusing on health information has increased due to the greater significance of health information retrievals by information seekers. The Text REtrieval Conference (TREC) focused on addressing questions that patients may come up with after reading their medical reports [3]. This task concentrated on a particular topic about a specific disease and relative treatment. All positive contributions made by participants on how to improve the existing retrieval systems were submitted and deliberated [8].

2.1 Information Retrieval Approaches

This section discusses numerous approaches deployed in addressing the problems of user-centered health information retrieval. The study classifies the information retrieval approaches under three fundamental modules; Language Modeling (LM) Models approach, Vector Space Models approach (VSM) and Context Based Approach (CBA).

2.1.1 Language Modeling (LM) Model Approach

Language Modeling Model approach is an arithmetical distribution model that allocate likelihoods to an order of terms, which predicts the possibility of their appearance within the script. All language based built systems operates on likelihoods for each term that come across and these likelihoods are self-determined on the nature of text. Various researchers used Language Modeling model for information retrieval system [2, 7, 9–11].

Shen [2] used language modelling approach and Indri search engine platform as their baseline with integrating Dirichlet smoothing. For the purpose of query expansion, the external resources Unified Medical Language System (UMLS) and Metamap were integrated. Moreover, related information from the user query logs were used to perform query expansion and the connection of shared words from query logs alongside medical vocabulary were used for information retrieval using concept-based approach. This method could not address the readability and vocabulary mismatched issues encountered by laymen in exploring extracted information online. However, this model achieved 0.7560 and 0.4016 for P@10 value and MAP value respectively. Moreso, Oh et al. [10] used Lucene search engine platform as their baseline and proposed the use of multiple-stage re-ranking method. This work is similar with [2] in such a way that they also used Dirichlet smoothing. Medical terms were extracted from discharge summaries to launch query expansion. This method obtained MAP value of 0.3989 and P@10 value of 0.74.

Work by Choi and Choi [7] is similar with [2, 7, 10] where they used indri as search engine with language model and Dirichlet smoothing correspondingly. They combined query expansion using the Metamap vocabulary, with discharge summaries and initial query extracted terms coordinated together. They conducted experiment with the superior feature by integrating learning to rank methods. This feature determines which of documents to be appeared and counted the frequency of terms prior-hand. This model managed to achieve 0.3494 of MAP score and 0.75 for P@10 score. Also Saleh [9] and Pecina [2] presented an interesting variant using Hiemstra language model with terrier search engine. The performance of the system was improved by using pseudo random feedback and Medline resource during query expansion. HTML strip and Boiler pipe resources were incorporated to decrease the data size to 6% of the original by removing insignificant terms. The system overall performance conveyed a MAP value of 0.1677 and a P@10 value of 0.5360.

2.1.2 Vector Space Model (VSM) Approach

Vector space model represents bits of script as vectors of identifiers. It was first tested in the SMART information retrieval system [4]. There are several variants have been offered such as, generalized vector space model and weighted vector space model. The vector space model based systems are easier to operate due to the fact that they are linear algebra oriented, it calculates the degree of resemblance between huge documents and queries with a limited similarity supports. However, it is deficient when representing long text documents by showing very poor similarity values.

Ksentini et al. [7] used the vector space model for the information retrieval approach. The cosine degree was used to measure the similarity between the query and document. The link between a query term and the total group of documents are computed by increasing the weighted Term Frequency (TF) and Inverse Document Frequency (IDF) measures. The default setup Terrier retrieval system is used for discontinue term deletion, tokenization and lessening. The overall scheme performance attained a MAP score of 0.167 and a P@10 score of 0.55.

Thesprasith and Jaruskulchai [6] used the Lucene search engine as their baseline system for stemming and tokenization. Pseudo-relevance feedback method was incorporated for the purpose of query expansion. The concepts mined from Medline biomedical dictionary were appended to expand the search query. Rocchio's formula was the determinant factor for extracting terms and was used for SMART system stoppage, with routine assessment of Pseudo-relevance response. The system managed to achieve the MAP value of 0.20 and P@10 score of 0.5540. Also Ozturkmenoglu et al. [5] used terrier search engine as their baseline engine. Unlike work by Thesprasith and Jaruskulchai [6], the query expansion for a specific query was determined by the integrated probabilistic Naïve Bayes. The overall method performance was stated to achieve P@10 value of 0.67 and a MAP score of 0.305.

2.1.3 Context Based Approach (CBA) Approach

The Context Based Approach attempt to extend each query by first extracting all the concepts terms from the search query and the synonyms of these general concepts terms from the incorporated vocabularies e.g. UMLS in order to launch an expanded search query. It uses search engine platforms such as Indri, Lucene etc. as the baseline engines and UMLS, HTML and Medline as incorporated external resources for the purpose of query [12, 13].

Salton [12] used the context based approach with terrier search engine. The performance of the system was improved by using pseudo random feedback and Medline resource during query expansion. MeSH, Metamap resources were incorporated for the purpose of query expansion. The entire system performance experienced a great effect due to the presence of these resources. The system overall performance conveyed a MAP value of 0.1732, P@10 value of 0.5512 and NDCG@10 value of 0.5211.

Work by Suominen et al. [14] used the context based approach with indri as search engine as their base line system. They combined query expansion using the Metamap vocabulary, with discharge summaries and initial query extracted terms coordinated together. They conducted experiment with the superior feature by integrating UMLS and Metamap external resources. This feature determines which of the documents to appear and counted the frequency of terms prior-hand. This model managed to achieve 0.1732 of MAP score, 0.6122 for P@10 score and 0.5523 of NDCG@10 score. Also Koopman et al. [13] used terrier search engine as their baseline engine. Unlike work by

Suominen et al. [14] the query expansion for a specific query was determined by the integrated probabilistic Naïve Bayes. The overall method performance was stated to achieve P@10 value of 0.5324, MAP score of 0.3244 and NDCG@10 score of 0.5788.

3 Information Retrieval Approaches

This paper performs a comparative study on various approaches deployed to address the problem of user centered health information retrievals as shown in Table 1. The problems encountered specifically by laymen patients and the care giver include the readability and vocabulary mismatched issues in exploring information extracted from medical discharge documents and clinical reports online.

3.1 Dataset

The dataset used in this research work were provided by the Unified Medical Language System (UMLS), Medical Subject Heading (MeSH), Metamap and Khresmoi project6 [1]. These datasets covers a wide range of patients' information and medical topics. All documents in the collection are downloaded from numerous online sources, including Health on the Net organization certified websites, Genetics Home Reference, Clinical.gov and Diagnosia7 [12].

3.2 Differences Among the Three Information Retrieval Approaches (LM, VSM and CBA)

The major difference among the three approaches deployed in addressing problems of user-centered health information retrieval systems used in these comparative studies is that, The vector space model based systems are easier to operate due to the fact that they are linear algebra oriented, it calculates the degree of resemblance between huge documents and queries with a limited similarity supports. However, it is deficient when representing long text documents by showing very poor similarity values [3, 4]. The Language Modeling Model approach is an arithmetical distribution model that allocates likelihoods to an order of terms, which predicts the possibility of their appearance within the script. All language based built systems operates on likelihoods for each term that come across and these likelihoods are self-determined on the nature of text. Various researchers used Language Modeling model for information retrieval system [2, 10]. While The Context Based Approach attempt to extend each query by first extracting all the concepts terms from the search query and the synonyms of these general concepts terms from the incorporated vocabularies e.g. UMLS in order to launch an expanded search query. It uses search engine platforms such as Indri, Lucene etc. as the baseline engines and UMLS, HTML and Medline as incorporated external resources for the purpose of query [13, 14].

Table 1. Comparative study of system performances of LM, VSM and CBA approaches

	Authors	Approach	Datasets and query expansion used	Performances of the systems		
				P@10	NDCG@10	MAP
1	Shen et al. [2]	LM	The dataset used were Metamap, UMLS and the Query Expansion used were also UMLS, Mutal Information and Metamap	**0.7560**	**0.7445**	0.4016
2	Oh and Jung [10]	LM	The dataset used were None and the Query Expansion used were Abrv. + Pseudo random feedback	0.74	0.73	0.3989
3	Claveau et al. [9]	LM	The dataset used were Ogmios NLP, Metam TreeTagger, UMLS, FASTR, YATEA, and the Query Expansion used were FASTR morpho-syntactic variants, UMLS synonyms and abbreviations	0.6740	0.6793	0.4021
4	Thakkar et al. [1]	LM	The dataset used were Metamap, MeSH and the Query Expansion used were Pseudo relevance feedback and Query-likelihood	0.7060	0.6869	**0.4146**
5	Choi and Choi [7]	LM	The dataset used were Metamap, UMLS and the Query Expansion used were Discharge summaries and Intersection of terms from query	0.75	0.70	0.3494
6	Yang et al. [8]	LM	The dataset used were Metamap and the Query Expansion used were Pseudo relevance feedback and Markov random field f	0.69	0.6705	0.3589
7	Thesprasith and Jaruskulchai [6]	VSM	The dataset used were Medline and the Query Expansion used were Pseudo relevance feedback	0.5540	0.5471	0.2076
8	Ozturkmenoglu et al. [5]	VSM	The dataset used were Medline and the Query Expansion used were Naïve bayes probabilistic expansion	0.6740	0.6518	0.3049
9	Drame et al. [4]	VSM	The dataset used were MeSH Metamap, UMLS and the Query Expansion used were UMLS Synonyms	0.5460	0.5574	0.2315
10	Ksentini et al. [3]	VSM	The dataset used were None and the Query Expansion used were Weighted vectors for query terms	0.5460	0.5625	0.1677
11	Koopman et al. [13]	CBA	The dataset used were Medline and the Query Expansion used were Naïve bayes probabilistic expansion and Terriers	0.5324	0.5788	0.3244
12	Suominen et al. [14]	CBA	The dataset used were Metamap, UMLS and the Query Expansion used were UMLS Synonyms	0.6122	0.5523	0.2011
13	Salton [12]	CBA	The dataset used were Metamap, MeSH and the Query Expansion used were Pseudo relevance feedback	0.5512	0.5211	0.1732

3.3 Performance Metrics

The Health Information Retrieval approaches will be evaluated using three parameters that are:

(a) *Precision at 10 documents (P@10)*
P@10 calculates the proportion of relevant documents at every 10 documents retrieved from a query. It can be computed as $P@10 = \frac{(A)10}{(A+B)10}$ where P is a Proportion of Relevant Documents Retrieved at every 10 document, A is Retrieved Relevant Documents and B is Retrieved Non relevant Documents [11].

(b) *Normalized Discounted Cumulative Gain at 10 documents (NDCG@10)*
NDCG@10 computes the cumulative gain at each position for a chosen value of p for the entire relevant document in the query. NDCG@10 is computed as $NCDG_p = \frac{DCG_p}{IDCG_p}$ where $IDCG_p = \left(IDCG_p\right) = \sum_{i=1}^{REL} \cdot \frac{2^{reli}-1}{IDCGp}$. REL represent the list of relevant documents, DCG_p is used to emphasize highly relevant documents appearing early in the result list [11].

(c) *Mean Average Precision (MAP)*
MAP computes the Mean Average Precision of relevant documents retrieved from a query. MAP is computed as $MAP = \frac{1}{N}\sum_{J=1}^{N} \cdot \frac{1}{Qj}\sum_{i=1}^{Qj} \cdot P(doc_i)$ where, Q_j is number of relevant documents for query j, N is number of queries and $P(doc_i)$ is precision value at ith relevant document [11].

4 Result and Discussions

For the specific task of comparing the various approaches deployed by previous researchers in addressing the problems of user centered health information retrievals and identifying the better approach to be deployed by researchers in addressing the aforementioned problems. The Language Based approach focused on addressing readability issues encountered by laymen patients and their relatives in exploring information extracted from medical discharge documents and clinical reports online; vocabulary mismatched issues between laymen queries and expert vocabulary during query expansion which affects the information retrieval system performance. The outcome of the study is as presented in Table 1 and Fig. 1.

Work by [8, 2] used the Language Model approaches in extracting medical concepts from the search queries and synonym terms from incorporated external resources. Only most specific medical concepts in the layman search queries were extracted and their synonym terms from incorporated external resources were expanded into a new query. The outcome of the search results contain less medical concepts which clearly attempted to better address the readability and vocabulary mismatched issues. There is an increased understanding of retrieved results due to less medical concepts in the returned results. The state-of-art results were reported by [2] using Language Model Based Retrieval System with a P@10 of 0.7560 and a NDCG@10 of 0.7445.

Works by [4–6, 7] used the Vector Space Models approaches and [12–14] used the Context Based Approaches in extracting medical concepts from the laymen search queries and synonym terms from incorporated external resources. The induced search results returns information with more medical concepts compared to the Language Model approaches. This shows that the Vector Space Model and Context Based Approaches concentrates less on most specific terms during the query search and also fail to incorporate controlled vocabulary during query expansion which affects their results.

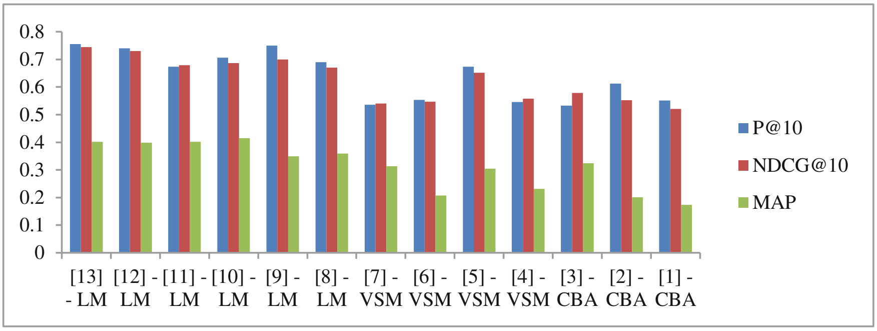

Fig. 1. Comparison of various methods based on and P@10, NDCG@10 and MAP values

The scientific reasons for the results obtained by LM, VSM and CBA could be explained in reference to this query "Anoxic Brain Damage" with label number (abd2017001), only the most specific term "Anoxic Brain Damage" and its synonyms "Anoxic Encephalopathy" and "Anoxic Brain Injury", are expanded into the new query by the language modeling model (LM), thereby disregarding the general terms Brain, Brain Injury, and Injury unlike VSM and CBA that extracts both the general and most specific terms into the new query thereby causing readability and vocabulary mismatched issues in exploring information extracted from medical discharge documents and clinical reports. In the same vain for this second query, "Stroke and Respiratory Failure" with label number (abd2017002), is made up of two labeled most specific terms "Stroke" and "Respiratory Failure". The extended query will comprise of "Cerebrovascular Accident", "Vascular Accident, Brain", "Kidney Failure", and Renal Failure.

Based on the comparative studies, the best results were achieved by [2] with P@10 value of 0.7560 and NDCG@10 value of 0.7445. Moreover, work by Thakkar et al. [1] presented the highest value of MAP with 0.4146. This work was based on language modeling methods and query expansion performed on specific medical concepts and synonym terms extracted from the search query. They also incorporated two controlled vocabularies that are Metamap and UMLS for the specific purpose of addressing vocabulary mismatched issue.

4.1 Statistical Test for the Validation of Experimental Results

A Language Model (LM) is a statistical distribution model that assigns probabilities to terms in the sequence which predicts the possibilities of such term for appear in the document. The model defines a probability P(T/D), where T refers to terms and D represents Documents. From my experimental results, since the process is repeated 10 times for each performance metrics P@10, MAP and NDCG@10, picking any relevant term in the document one at a time T_1, T_2, $\ldots T_{n=10}$ in D is given by P(T_1, T_2, … $T_{n=10}$ = $\sum_{i=1}^{n=10} P(Ti/D)$. The equation will assign zero (0) probability to all irrelevant terms and one (1) to relevant terms in the sequence. The statistical analysis show that out of every10 terms retrieved from a document (i.e. 100% of the terms), Language Modeling Model (LM) approximately scored 0.76 (76%) as relevant and readable retrieved documents in respect to P@10, as against Vector Space Model (VSM) 0.67 (67%) and Context Based Approach (CBA) 0.57 (57%). In respect to NDCG@10, LM scored 0.74 (74%), VSM 0.65 (65%) and CBA 0.57 (57%) as relevant and readable retrieved documents and finally in relation to MAP, LM scored 0.41 (41%), VSM scored 0.30 (30%) and CBA scored 0.32 (32%).

5 Conclusion

Based on the outcome of the experimental results obtained from Fig. 1 clearly shows that out of every 10 documents retrieved from a medical search result (i.e. 100% of the documents), Language Modeling Model (LM) approximately scored 0.76 (76%) as relevant and readable retrieved documents in respect to P@10, as against Vector Space Model (VSM) 0.67 (67%) and Context Based Approach (CBA) 0.57 (57%). In respect to NDCG@10, LM scored 0.74 (74%), VSM 0.65 (65%) and CBA 0.57 (57%) as relevant and readable retrieved documents and finally in relation to MAP, LM scored 0.41 (41%), VSM scored 0.30 (30%) and CBA scored 0.32 (32%).

The experimental results obtained in Fig. 1 shows that the Language Modeling is the best approach to be used in addressing information retrieval system problems. Every 76% of retrieved information by Language Modeling Model are Readable and Vocabulary Mismatched free as against Vector Space Model (VSM) 67% and Context Based Approach (CBA) 57%. The Language Modeling approach overcomes the vector space model and Context Based Approaches between 3 and 8% in relation to Precision@10 and 20% in MAP value. These better results were achieved due to Language Modeling Approaches fully concentrating on most specific terms during query search rather than general terms and also incorporating controlled vocabulary such as Metamap and UMLS during query expansion. We recommend further work on this study should include the design of algorithm that will address the readability and vocabulary mismatched issues encountered from retrieved images, audios and videos.

Acknowledgements. The authors express appreciation to the Universiti Tun Hussein Onn Malaysia (UTHM). This research is supported by Short Term Grant vot number U653 and Gates IT Solution Sdn. Bhd. under its publication scheme.

References

1. Thakkar, H., Iyer, G., Shah, K., Majumder, P.: Team IRLabDAIICT at ShARe/CLEF eHealth 2014 task 3: user-centered information retrieval system for clinical documents. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
2. Shen, W., Nie, J.Y., Liu, X., Liui, X.: An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM @ CLEF2014eHealth task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
3. Ksentini, N., Tmar, M., Gargouri, F.: Miracl at CLEF 2014: eHealth information retrieval task. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
4. Drame, K., Mougin, F., Diallo, G.: Query expansion using external resources for improving information retrieval in the biomedical domain. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
5. Ozturkmenoglu, O., Alpkocak, A., Kilinc, D.: Demir at CLEF eHealth: the effects of selective query expansion to information retrieval. In: proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
6. Thesprasith, O., Jaruskulchai, C.: Csku gprf-qe for medical topic web retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
7. Choi, S., Choi, J.: Exploring effective information retrieval technique for the medical web documents: SNUMedinfo at CLEFeHealth2014 task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
8. Yang, C., Bhattacharya, S., Srinivasan, P.: The University of Iowa at CLEF 2014: eHealth task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
9. Claveau, V., Hamon, T., Grabar, N.,, Maguer, S.L.: RePaLi participation to CLEF eHealth IR challenge 2014: leveraging term variation. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
10. Oh, H.S., Jung, Y.: A multiple-stage approach to re-ranking clinical documents. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
11. Dybkjaer, L., Hemsen, H., Minker, W.: An overview of evaluation methods. In: Evaluation of Text and Speech Systems in TREC Ad-hoc Information Retrieval and TREC Question Answering. Springer, Dordrecht, the Netherland (2015)
12. Salton, G.: The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall Inc, Upper Saddle River, NJ, USA (2015)
13. Koopman, B., Zuccon, G., Bruza, P., Sitb on, L., Lawley, M.: An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In: Proceedings of CIKM (2012)
14. Suominen, H., et al.: The Proceedings of the CLEFeHealth2012—the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis. NICTA (2015)