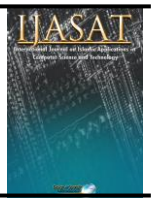


Content Management System On Prophetic Food with Multilingual Resources

Norita Md Norwawi[1], Muhammad Widus Sempo[2], Sundresan Perumal[3], Emran Huda[4]

[1,3,4]Faculty of Science and Technology, Universiti Sains Islam Malaysia, Malaysia
[2]Faculty of Quranic and Sunnah Studies, Universiti Sains Islam Malaysia, Malaysia
[1]norita@usim.edu.my,[2]widus81@usim.edu.my,
[3]sundresan.p@usim.edu.my,[4]emran.huda@usim.edu.my

ABSTRACT
Content management system (CMS) manages web content according to a predesigned information and database template. The content of the resources varies and may be in different languages. The different linguistic nature leads to language ambiguity and semantic of the content. Thus, the diversity of perspectives and context in handling heterogeneous knowledge resources need to be addressed. This paper presents a CMS that manages a one-stop knowledge repository on prophetic food from various validated resources such as al-Quran and Hadith which are in Arabic, classical manuscript in Malay language either written in Roman or Jawi text and scientific articles which are in English. The knowledge sources were verified by domain experts. A semantic query dynamism was the key feature proposed in the CMS implemented as a proof of concept for the one-stop knowledge repository on prophetic food called Naqli Aqli Semantic Search Engine (NAISSE). The offline and online translation demonstrated the mechanism to match the query with the multilingual documents.

Keywords: content management system, cross-lingual information retrieval, semantic search

1.0 Introduction

Handling heterogeneous knowledge sources is challenging due to the diversity of perspectives and context of the knowledge including knowledge elicitation. Knowledge identification and encoding, elicited from heterogeneous sources into a form understood by a computer involve a complex task (Aldhlan et al., 2010). The knowledge from the multiple sources needs to be modelled across the similarities and differences highlighting the major concepts and high relational structure among these concepts. In this study, one of the key differences includes the original language of the resources.

Due to the advancement of the Internet technology, an abundance of information are available on the World Wide Web (WWW) across the boundary of language. However, today due to the diversity of information and language barriers, there is a need to overcome the barriers in communication and cultural interchange across the globe.

An online information retrieval (IR) application provide access to multiple machine-readable resources easily and rapidly. There are 4 categories of retrieval which are (i) mono-lingual (ii) bi-lingual (iii) cross-lingual and (iv) multi-lingual (Agebele et al., 2018). Cross-language and multi-lingual IR allow user queries in one language to retrieve documents in another language. The query is posed in the source language where the target language is in the

relevant document matched. Cross-lingual access opens the boundary of language and culture due to the rapid progress of Internet technology. Thus, translation is the key signature in the query processing task. According to Agbele et al. (2018), there are three translation approaches which are (a) document (b) query or (c) hybrid query document translation.

However, due to the different linguistic nature, the translation may misinterpret the user's true intention due to language ambiguity. Hence, ensuring the right context for the translation is a challenge (Abusalah et al., 2005; Aldhlan et al., 2010; Sharmal and Morwal, 2015; Elayeb and Bournas 2016). Jena and Rautaray (2019) highlighted that among the three sources of knowledge for query translation: dictionary, corpora or machine learning, the corpora-based approach has a lower ambiguity with moderate development cost, therefore has been adopted in this study.

Searching and retrieval become more complex for multilingual content due to the linguistic nature of grammar, synonym, context and translation. For example, in the Arabic language, date palm may mean نخل, رطب, تمر lead to uncertain in its true context and semantic. *Kurma* and *tamar* are the translation in the Malay language whereas *phoenix dactylifera* is the scientific name. This vagueness is due to homonymy and polysemy whereby one word may have two meanings depending on the context. Thus, matching with the right context need to be addressed properly. (Abusalah et al., 2005; Sharmal and Morwal,2015; Elayeb and Bournas 2016).

In this paper, the mechanism on handling the intricacies due to the multi-linguistic ambiguities in the CMS will be discussed. The corpora-based translation on query processing for multilingual CMS will also be presented including the implementation. The CMS complements the knowledge repository on prophetic food for maintaining and authenticating the content in general. Section 2 will present the methodology adopted in designing the proposed solution and Section 3 discusses the results from this work and concluded in Section 4.

2.0 Methodology

To resolve the cross-lingual issue in the IR, the query processing mechanism needs to be analyzed based on the resources used for the study. Offline corpora were developed that translate the original documents into the other two languages where necessary. Domain experts identify keywords related to a chosen prophetic food, the linguistic requirements and the connection to the information in the resources. The structure of the tables for the CMS and mapping to all relevant resources were analyzed and designed into the query processing mechanism. NAISSE CMS was developed as template use for updating verified content and maintenance purposes.

2.1 Multilingual Resources on Prophetic Food

A one-stop knowledge repository on prophetic food with an IR capability was the main purpose of this research, However, since the resources are in classical Arabic for the Quran and Hadith, Malay language either in Roman or Jawi text for classical manuscripts and English for the scientific article in general, a cross-lingual query processing is needed. A content management system was developed to store the machine-readable form of the resources using a knowledge management approach. Two key elements of the IR which are query processing and indexing need to cater the multilingual nature of the resources. Fig. 1 shows the multilingual nature of the one-stop knowledge repository on prophetic food called Naqli Aqli Semantic Search Engine (NAISSE).

Fig 1. Multi-lingual nature of NAISSE: One-Stop Knowledge Respository on Prophetic Food related to Dates and Goat's Milk.

2.2 Systematic Review and Content Analysis of Resources.

Prophetic food was the focus of this study naturally has an important link to the al-Quran and Sunnah besides classical manuscripts on traditional remedies and medication for health. On the other hand, the scientific study confirms the healthy sustainable prophetic wisdom and lifestyle of Rasulullah SAW with empirical evidence. Team members with expertise from various professional backgrounds such as Quran and Sunnah studies, manuscript, food technologist and medical conducted the systematic review and content analysis related to dates and goat's milk, the scope chosen for the study.

2.3 Knowledge Mapping and Keywords Index

The information collected from the systematic review was tabled out and mapped using keywords identified by domain experts. Information includes details on Quranic verses, Hadith sources, manuscripts details including a snapshot of the section on the page and URL links to scientific articles and abstract. The translation in two other languages, synonyms and context meaning of the keywords were also selected. Fig. 2 depicted the mapping to the resources and linguistics aspect.

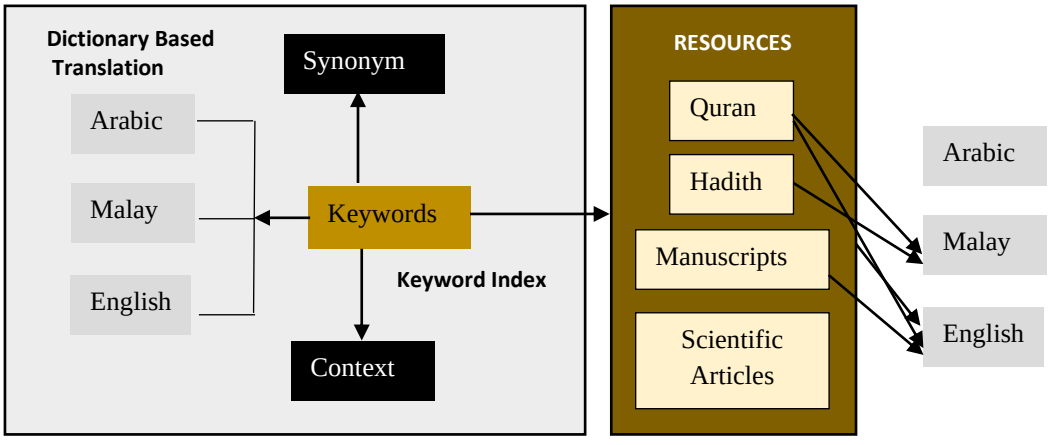


Fig. 2: Mapping between keywords to the multilingual resources and its linguistics aspect.

Fig. 3 illustrates the structure of the two tables created in the CMS

a. link keywords to the resources
b. link keywords to its translation, synonym and semantic.

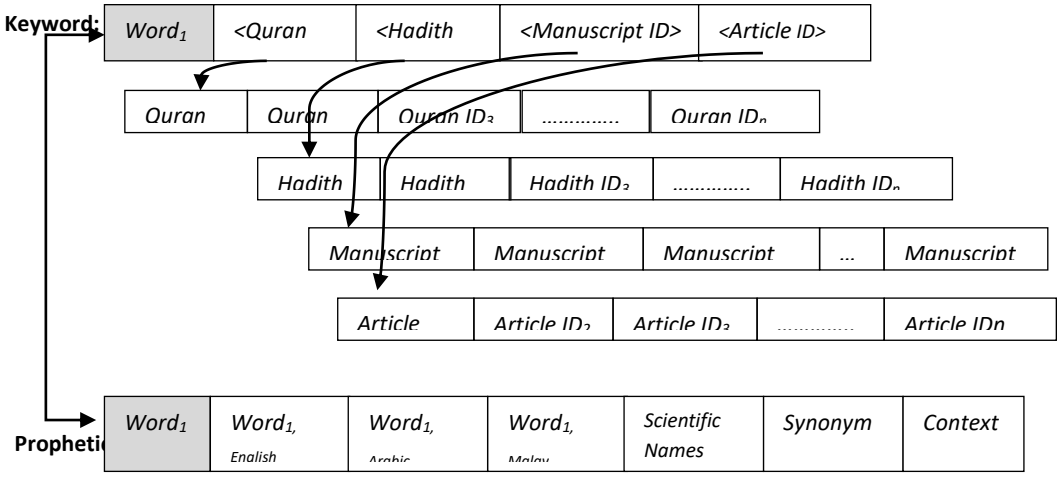


Fig. 3: Illustration of the CMS table structure linked to resources and linguistics expansion

2.4 Corpora Preparation

The knowledge repository consists of four resources: al-Quran, Hadith, classical manuscript in Roman and Jawi scripts and the scientific articles. An offline translation exercise was conducted for Al-Quran and Hadith into English and Malay translation recommended by domain experts. The classical manuscript in Malay was also translated into English and the Jawi text were rewritten in Jawi font and verified by the experts.

2.5 Query Processing Technique

The corpora based offline and dictionary-based online translation were integrated with the contextual meaning based on synonym and semantic. The query will be matched to keywords mapped to other languages available and searched the target resources accordingly. Fig. 4 shows the schematic diagram on the query dynamism.

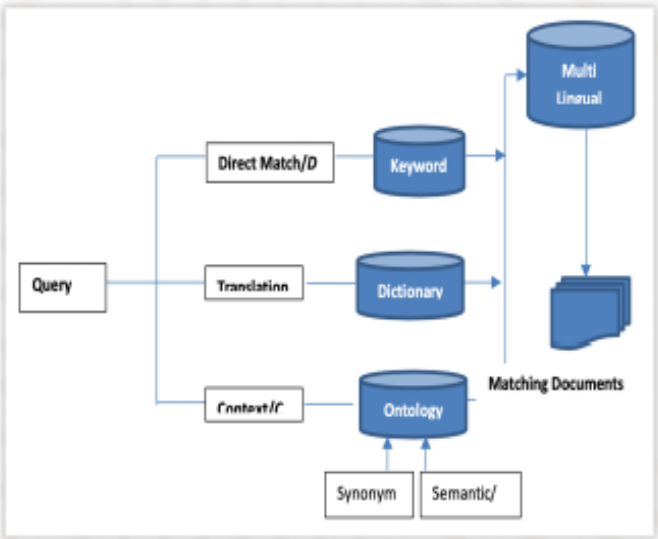


Fig 4. Schematic diagram of the three mechanisms of the query dynamism

Naqli Aqli Semantic Search Engine (NAISSE) is a web application developed adopting the search and retrieval of a query in a single keyword that may be in any of the languages. There are three major mechanisms in processing the query to find its relevant matching documents:

a. Direct matching of the keyword (***D***) used a list of predefined keywords by the domain experts.
b. Translation of the keyword into the other two languages (***T***) using a dictionary-based approach.
c. Word in context with the keyword (***C***) either through the synonym (***S***) or meaning (***M***) that semantically connected by defining the ontology of the prophetic food in relations to the resources.

For this study, dates and goat's milk was chosen as a test case for the proof of concept. For example, if the keyword is <*date*s>, through direct matching (***D***), any resources with the exact word <*dates*> will be matched and retrieved. Through its translation (***T***) into Malay and Arabic, new words listing will be generated eg. { kurma, تمر }. Finally, additional words will be appended to the list such as from the synonym eg. { kurma, tamar, تمر, نخل, رطب }. Then, this list of identified words will be used to match words in the documents if exist. This steps are represented using set theory as in Fig. 5.

Assume $\boldsymbol{L}$ is the set of all matching words, $\boldsymbol{w}$, from all the resources Quran ($\boldsymbol{Q}$), Hadith ($\boldsymbol{H}$), Malay manuscript in roman text ($\boldsymbol{S_r}$), Malay manuscript in jawi text ($\boldsymbol{S_j}$) and scientific articles ($\boldsymbol{J}$). Thus,

$$L = \{w| w \in \mathrm{D}\} \cup \{w| w \in T\} \cup \{w| w \in C\}$$

where $\boldsymbol{L}$ are the list of all matched keywords such that

$$D_a = \{w| w \in \mathrm{Q}\} \cup \{w| w \in H\} \quad \text{for Arabic text.}$$

$$D_m = \{w| w \in \mathrm{S_r}\} \cup \{w| w \in S_j\} \quad \text{for Roman or Jawi Malay text.}$$

$$D_e = \{w| w \in J\} \quad \text{for English text.}$$

thus, $\boldsymbol{D} = D_a \oplus D_m \oplus D_e$ implies only exact match in the same language as query word.

As for the dictionary-based translation $\boldsymbol{T}$,

$$T_m = \{w| w \in \mathrm{Q}\} \cup \{w| w \in H\} \cup \{w| w \in J\} \quad \text{for query in Malay language, } \boldsymbol{m}$$

$$T_a = \{w| w \in \mathrm{S_r}\} \cup \{w| w \in S_j\} \cup \{w| w \in J\} \quad \text{for query in Arabic language, } \boldsymbol{a}$$

$$T_e = \{w| w \in \mathrm{Q}\} \cup \{w| w \in H\} \cup \{w| w \in \mathrm{S_r}\} \cup \{w| w \in S_j\} \quad \text{for query in English language, } \boldsymbol{e}$$

and

$$\boldsymbol{T} = \{w_l | w_l \in T_a\} \cup \{w_l | w_l \in T_m\} \ \& \ l \neq a, l \neq m \oplus$$

$$\{w_l | w_l \in T_a\} \cup \{w_l | w_l \in T_e\} \ l \neq a, l \neq e \oplus$$

$$\{w_l | w_l \in T_m\} \cup \{w_l | w_l \in T_e\} \ l \neq m, l \neq e$$

where is $\boldsymbol{T}$ is the set of translation of the keyword $\boldsymbol{w_l}$ in other language

Finally, the context based match $\boldsymbol{C} = \{w_s| w_s \in N\} \cup \{w_m| w_m \in O\}$
where w_s are words matched in the synonym list, $\boldsymbol{N}$ and w_m are words matched from the ontology O.

Fig. 5: A Semantic Query Dynamism Set Representation for Cross-Lingual Information Retrieval

3.0 Results and Discussion

NAISSE is developed to search reliable and original resources on prophetic foods from the multilingual resources mentioned. The NAISSE system provided two platforms for different end-users. The first platform, which is in **naisse.org**, is the main web application provided for the subscribing users to access information that they need on prophetic foods. The second platform, which is the administrator panel for the source provider to dispense related information, in comprehensive and thorough details. NAISSE app was also developed for both IOS and Android operating system.

NAISSE content is managed by the multilingual capability CMS developed where the users are the domain experts who will upload new content accordingly thus ensuring verified and authentic knowledge. Fig. 5 shows the CMS dashboard that summarizes the information in the repository such the number of keywords recorded, the number of users that have subscribed to the NAISSE website, as well as the number of prophetic foods was recorded in the system. The administrator will also be to view the last query searched by the users and the IP addresses. S/He may view and manage the keywords related to the Quran and hadith, manuscript, scientific articles on prophetic food identified by domain experts. Fig. 6 is a snapshot of the dashboard.

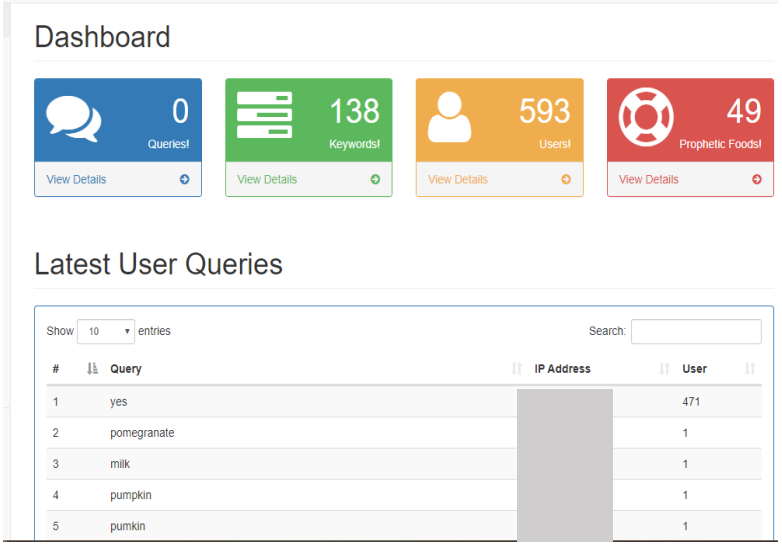


Fig. 6: Screenshot of the NAISSE CMS dashboard

NAISSE CMS was designed to have connected components defined by the keywords, its translations, synonyms and meaning in context. For example, a wedding is semantically connected to dates in some hadith being the food for the wedding feast. NAISSE CMS provide a template for the domain expert to create the connection of the knowledge through the keyword. Thus, S/He must also provide the translation and the synonym of the keyword identified based on their interpretation from the Quran and Hadith being the primary sources of knowledge in Islam and link it to other knowledge such as on food or medical sciences. Fig. 7 illustrates the connection of an identified keyword to verses of the Quran, Hadith text, manuscript text and scientific articles.

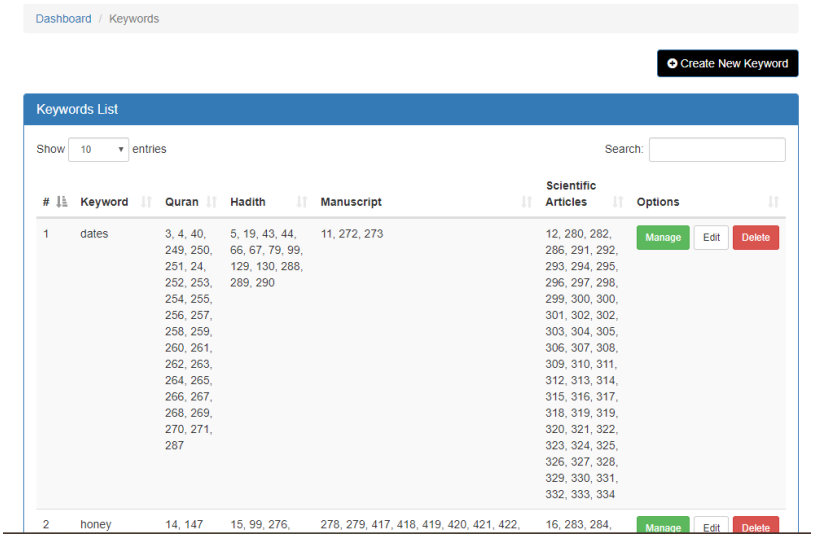


Fig. 7: Screenshot of keywords link to various resources ID identified by the domain expert.

The administrator and domain expert who is given the authority to access NAISSE CMS will be able to edit, insert additional information, new keywords and its relevant verified information related to Quran, hadith, manuscript, and scientific articles. As for the manuscript's page, the admin may add new items from an identified manuscript, commonly in two languages (Malay, English) with manuscript bibliography details and the image for the proof of the manuscript. The manuscript which is in Jawi will be reproduce using Jawi font and roman text verified by experts.

The query dynamism in Fig. 4 and 5 was mapped into a relationship based on the domain expert's analysis and interpretation to protect the integrity of the knowledge repository contains in the specialized search engine with predefined databases. Full automation of keyword matching with auto-translation may be risky on the authenticity aspect of the knowledge. Quranic verses and Hadith text are considered as sensitive data that requires stringent content integrity features. Fig. 8 shows an initial work in translating the schema of the query dynamism embedded in the CMS.

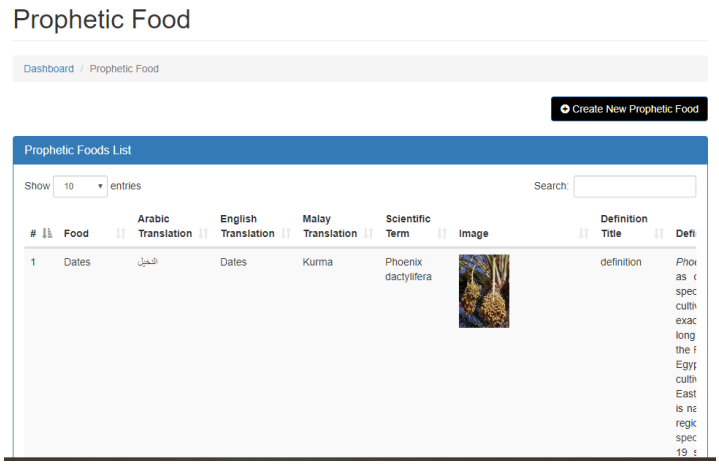
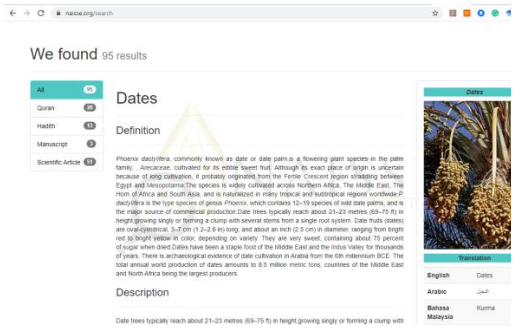
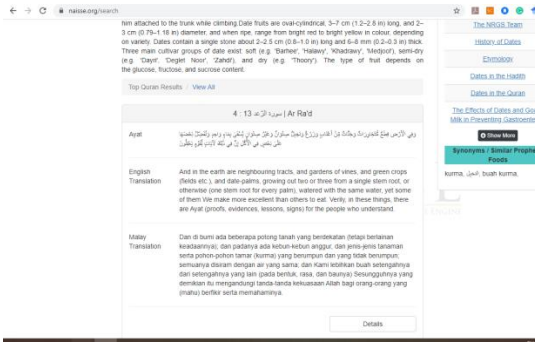


Fig. 8: Initial version of the semantic query dynamism implementation in NAISSE CMS involving direct matching (***D***) and translation mechanism (***T***).

Fig 9a,b,c and d shows the display of from NAISSE web application and NAISSE apps answering a query on dates that are mentioned in all the resources.



a. Screenshot of main display matching with keyword date



b. Screenshot of main display matching with keyword date that exhibits the synonym

c. Screenshot of matching manuscripts text with keyword date

d. Screenshot of matching scientific articles with keyword date

Fig 9. Screenshots from NAISSE web and apps application with keyword *dates*.

4. Conclusion

The CMS provides a template for the multilingual searching, matching and retrieving information related to a query in three major steps: direct match with current keywords, dictionary-based translation and context-based matching through synonyms or semantic aspect. Since the Quran and Hadith are sensitive data with high content integrity requirements, full automation may affect the authenticity of the information. The query dynamism determines the CMS design and architecture that reflects the three-mechanism defined. To further protect the integrity, the domain expert is responsible for identifying and verifying the content in NAISSE CMS. For future work, an automated keywords generation can be extracted from the corpora to avoid mismatched or unmatched query.

Acknowledgements

The authors express their appreciation for the Niche Research Grant Scheme, awarded by the Ministry of Education to Universiti Sains Islam Malaysia (USIM) with project code USIM/NRGS-P/FST/8404/52113.

References

Abusalah, M., Tait, J., Oakes, M. (2005). Literature Review of Cross-Language Information Retrieval. World Academy of Science, Engineering and Technology 4 175-177

Agbele, K., K., Ayetiran, E.F., Aruleba, K.D. (2018). Survey on Cross-Lingual Information Retrieval. International Journal of Scientific & Engineering Research, 9 (8): 484-491

Aldhlan, K. A. and Zeki, A. M. and Zeki, A. M. (2010) *Data mining and Islamic knowledge extraction: Alhadith as a knowledge resource.* In Proceedings of the International Conference on Information and Communication Technology for Muslim World (ICT4M.

Elayeb, B, Bournas. I. (2016). Arabic Cross-Language Information Retrieval: A Review. ACM Trans. Asian Low-Resou. Lang. Inf. Process Vol. 15, No. 3

Jena, G.C., Rautaray, S.S. (2019). A comprehensive survey on cross-language information retrieval system. Indonesian Journal of Electrical Engineering and Computer Science 14 (1): 127-13

Sharmal M, Sudha Morwal. (2015). A Survey on Cross-Language Information Retrieval. International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2. pp384-387

Biodata

	Norita Md Norwawi is a professor in Computer science at the Faculty of Science and Technology, Universiti Sains Islam Malaysia. She obtained her BSc in Computer Science at the Universiti of New South Wales, MSc in Computer Science at Universiti Kebangsaan Malaysia and PhD at Universiti Utara Malaysia. Her research interest is in Artificial Intelligence, Data Mining and Multiagent System.
	Muhammad Widus Ph. D is the researcher in the field of Quranic Study and Tafsir (exegesis), graduated from Al-Azhar University, Cairo, Egypt. The researcher has presented some works related to the issue of Quran and Tafsir in different conferences and seminars both national and international. He also has published some articles in journals both national and international. Besides, He authored some books on Quranic study and tafsir
	Sundresan Perumal is a senior lecturer, computer forensic investigator, network security consultant and noted authority on electronic evidence. He is currently heading the Innovation and Commercialization Unit in University Science Islam Malaysia. His research specialization are in Digital Forensic , Advanced Persistent Threat(APT), Internet of Things Forensic , and Internet of Medical Things (IoMT).
	Emran Huda is a masters (by research) student in Computer science at the Faculty of Science and Technology, Universiti Sains Islam Malaysia. He obtained his BSc in Computer Science and Software Engineering from Swinburne University of Technology. Currently he is pursuing his research, to Develop an Artificial Intelligence (AI) Algorithm Based E-counselling system.

Abstract in Arabic

نظام إدارة المحتوى الخاص بالطعام النبوي
مصادر متعددة اللغات

نوريتا محمد نورواوي[1] محمد ويدوس سيمبو [2] ساندرسان فرومال[3] عمران هدى[4]

[1]نوريتا محمد نورواوي، كلية العلوم والتكنولوجيا، جامعة العلوم الإسلامية الماليزية. Email: norita@usim.edu.my
[2]محمد ويدوس سيمبو، كلية القرآن والسنة، جامعة العلوم الإسلامية الماليزية.Email: widus81@usim.edu.my
[3]ساندرسان فرومال، كلية العلوم والتكنولوجيا، جامعة العلوم الإسلامية الماليزية. Email: sundresan.p@usim.edu.my
[4]عمران هدى، كلية العلوم والتكنولوجيا، جامعة العلوم الإسلامية الماليزية

الخلاصة. نظام إدارة المحتوى (Content management system) يقوم بتنظيم محتوى شبكة الانترنت (web) وفقًا لنموذج قاعدة بيانات ومعلومات مسبقة التصميم. ومحتوى الموارد يظهر بأشكال متنوعة وقد يكون بلغات مختلفة. وقد تؤدي الطبيعة اللغوية المختلفة إلى بعض الإشكالات في اللغة ودلالة المحتوى. ولهذا، يجب معالجة تنوع وجهات النظر والسياق في التعامل مع موارد المعرفة غير المتجانسة. تقدم هذه الورقة نظام إدارة المحتوى الذي يدير مخزن المعرفة فيما يتصل بالأغذية النبوية من مختلف الموارد العلمية مثل القرآن، والحديث ومخطوطات ملايوية، ومقالات علمية باللغة الإنجليزية. وقد تم التحقق في هذه المصادر من قبل خبراء المجال. وقد كانت دينامية الاستعلام الدلالية هي السمة الرئيسية المقترحة في CMS التي تم تنفيذها كدليل على المفهوم لمستودع المعرفة الشامل للأغذية النبوية المسمى بمحرك البحث الدلالي (Naqli Aqli Semantic Search Engine). وأظهرت الترجمة العادية دون الاستعانة بالإنترنت والترجمة عبر الإنترنت تطابقًا بشكل ملحوظ في الاستعلام مع المستندات متعددة اللغات.

الكلمات المفتاحية: نظام إدارة المحتوى، استرجاع المعلومات عبر اللغات المختلفة، محرك البحث الدلالي