உத்தமம்
INFITT
உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம்
International Forum for Information Technology in Tamil

Penn
University of Pennsylvania

மாநாட்டுக் கட்டுரைகள்
CONFERENCE PAPERS

TAMIL INTERNET 2011

தமிழ் இணையம் 2011

Kuralagam - Concept Relation based Search Engine for Thirukkural

Elanchezhiyan. K, T V Geetha, Ranjani Parthasarathi & Madhan Karky
Tamil Computing Lab (TaCoLa)
Department of Computer Science and Engineering
College of Engineering Guindy, Chennai – 600025
E-mail: chezhiyank@gmail.com, madhankarky@gmail.com

Abstract

Thirukkural is one of the most popular literatures in Tamil Language. Thirukkurals are being quoted in speeches, news articles, blogs, micro-blogs and has a very strong reach in the Internet. Various interpretations of Thirukkurals have been proposed by eminent Tamil scholars. This paper aims at presenting the world's first conceptual search framework for Thirukkural. The Framework uses CoReX [1]; a concept relation based indexing and presents a ranking model based on concept strength and popularity of Thirukkurals, obtained by a Thirukural statistic crawler. The search Framework is evaluated using Average Precision and Mean Average Precision (MAP) was found to be 0.83 compared to 0.52 with traditional keyword based search.

1. Introduction

Thirukkural is the one of the outstanding accomplishment of Tamil literature. It had been translated in many languages. Thirukkural has totally 133 chapters. These are classified in to Aram (Virtue), Porul (Wealth) and Kamam or Inbam (Love). Each chapter has ten Thirukkural; Thirukkural in the form of couplets illustrates various aspects of life. Most of the present day Thirukkural search engines are keyword-based and Bilingual Keyword-based. Thirukkural search engines are available for Tamil and English language. Those keyword-based search engines fail to satisfy the user requirements. For example, in keyword-based search user won't get the result for the common word "பணம் "(Money). For the reason that actual keyword "பணம் "was not used in the Thirukkural.

The Kuralagam is a Conceptual and bilingual Thirukkural search engine. It is designed to clear up the complication in the traditional Thirukkural Search engine using CoReX Frame work. CoReX is designed such that the documents retrieved through it are semantically relevant to the query. The data structure used by the CoReX helps in storing concepts of terms rather than storing just words, thereby retrieving Thirukkurals that are semantically relevant to the query. The main purpose of such indexing techniques is cross lingual information retrieval by an intermediate representation called the Universal Networking Language [2] (UNL). The Universal Networking Language is an electronic language for computers to express and exchange information. Kuralagam search system fetches Thirukkurals based on keywords, concepts and expanded query words using the Concept Based Query Expansion technique using CoReX Framework.

In this paper we propose Kuralagam, a Concept Relation Based Thirukkural Search. Kuralagam aims at understanding the Thirukkural and its meaning by indexing them based on concepts and their relations rather than indexing the keywords and their frequencies. The Kuralagam was implemented and tested with 1330 Thirukkurals with 4 explanations (Kalaingar Karunanidi, Mu Va, Soloman Poppaiya, and G.U.Pope). The search results were compared against traditional keyword based search for precision and relevance.

2. Background

CoReX [1] is a concept based semantic indexing technique used to index Universal Networking Language (UNL) expressions. CoReX retains the semantics captured by the UNL expressions. Since UNL expressions are stored as graphs, CoReX uses graph properties to index the UNL graphs. CoReX considers the out degree of each node and frequency of the same for indexing which helps in capturing the relations between the concepts in UNL expressions thereby retaining the semantics of the same. CoReX is simple and efficient and helps in retrieving documents which are semantically relevant to the query. The Thirukkural popularity score is computed by giving a Thirukkural sequentially to the web and finding its frequency distribution across the popular blogs, news articles, social nets etc.

3. Thirukkural Search Framework

Thirukkural search framework presented in figure1 can be divided into two major divisions, online and offline, in terms of the time of processing. This section describes the various components of the Thirukkural search in detail.

3.1 Offline Processing

The offline process comprises indexing Thirukkurals and their interpretations and crawling the web for usage of each Thirukkural.

3.1.1 Web Crawler

A Thirukural statistics crawler crawls the news and blog documents on the web to find the usage of each individual Thirukkural. The usage on internet is recorded for measuring the popularity score for each Thirukkural, which is explained in detail later.

3.1.2 En-conversion

Here a Thirukkural and its meaning are passed to a rule based system to identify the various concepts in the Thirukkural and the rules are used to identify one of the 44 UNL relations [2]. Enconversion [4] uses the Morphological Analyser [3] to recognize various morphological suffixes of a word and uses this information along with syntax and semantics to identify the relationship between concepts. UNL graphs are generated for every sentence constituent. The UNL graph is then sent to CoReX indexer along with information such as Thirukkural Number, positional index and original keyword, its frequency in the document etc.

3.1.3 Indexer

The Kuralagam Indexer is designed based on CoReX Techniques. The Indexer stores and manages the UNL graphs in two different indices. Concept only index (C index), and Concept-Relation-Concept

index (CRC index) are the two indices maintained by the indexer. The UNL graphs are stored in the indices by their concept for efficient retrieval.

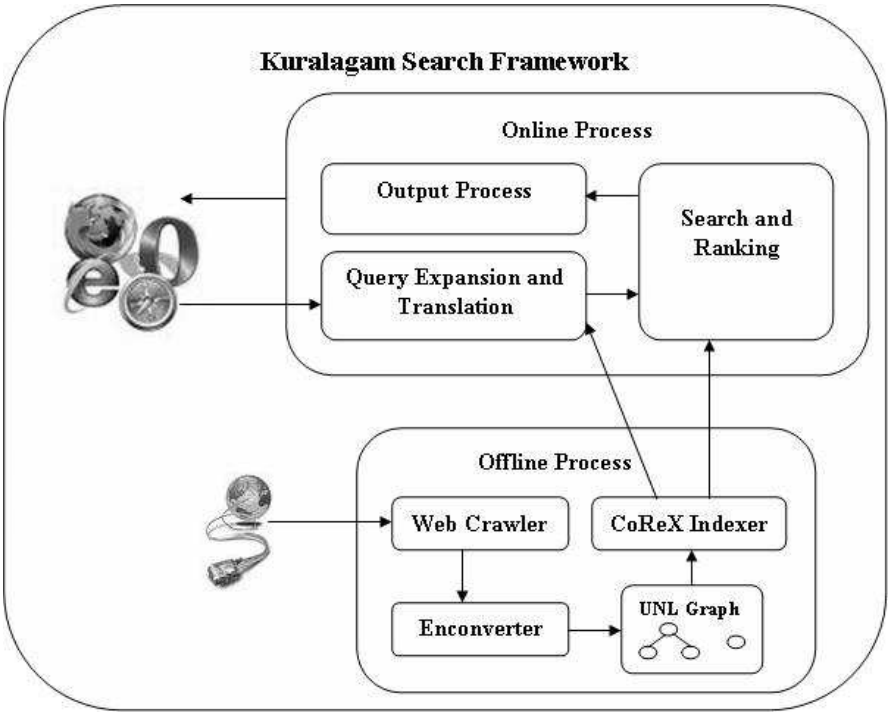

Fig 1: Kuralagam Search Framework

3.2 Online Processing

A user's query is processed, converted to UNL graph(s), expanded and sent to a search and ranking module, where the Thirukkurals that match the concept relation similarity are ranked and sent for output processing. The output processing module displays the retrieved Thirukkurals with its meaning and sends them to the user.

3.2.1 Query Translation and Expansion

A user query is first sent to Query Translation module. Query Translation module converts the user query to UNL graph. The Concepts in UNL graph are sent to the Query Expansion module. Query Expansion uses CRC (Concept Relation Concept) CoReX indices to fetch similarity thesaurus and co-occurrence list to populate the Multi list Data Structure.

3.2.2 Search and Ranking

The functionality of searching and ranking is to fetch the Thirukkural number and its details. Thirukkurals for a given query are fetched using the two types of concept relation indices namely CRC and C. The query concept is expanded using related CRC indices pointing to the query concept. This helps in retrieving many Thirukkurals conceptually related to the query. This kind of conceptual retrieval is highly impossible with key word Thirukkural search engines. The ranking is done by giving priority to the indices in the order CRC>C. The ranking is also based on the usage score and frequency occurrence of the query concept. Hence the search results are based not only on the query

term but also on the concepts related to the query term. The search results and performance analysis is discussed in the next section.

4. Kuralagam Search Results & Analysis

Kuralagam is a conceptual search framework for Thirukkural. Kuralagam, unlike traditional keyword based searches, identifies the concepts in a Thirukkuaral and their relationship with each other. All 1330 Thirukkurals and their meaning were crawled, enconverted and indexed for search.

4.1 Tabbed Layout

In this paper, we propose a Tabbed Layout for displaying the results to the user. The Tabbed layout shown in figure 2, displays the results in 3 tabbed boxes to a class of results based on the concepts and relationship between concepts. Figure 2 displays the results for the query நட்பின் சிறப்பு (*Natpin sirappu*). The first cell displays the results of the concepts that contain the actual keywords which are sorted by the relation they have between them. Second tab identifies results that contain concepts of actual keywords, relation between them and displays the results corresponding to நட்பு பெருமை (*Natpu Perumai*). The third cell is based on expansions of the query. Here results corresponding to நட்பு துன்பம் (*Natpu thunbam*), நட்பு கொள் (*natpu koL*), நல்ல நட்பு (*Natpu nalla*) etc are displayed. The snapshot presented in figure 2 is from our engine implemented from the CoReX framework.

4.2 Performance Evaluation

The accuracy of the Thirukkural search engine was measured using the Precision. Precision can be computed using the formula given below [5]. We compute the precision for the first 5, 10 and 20 Thirukkurals. The average precision and mean average precision for a set of queries will indicate the performance of the system.

Fig 2: Tab Layout

The comparisons between concept based search and keyword based search were measured using Average Precision methodology and the result is shown in figure 3.

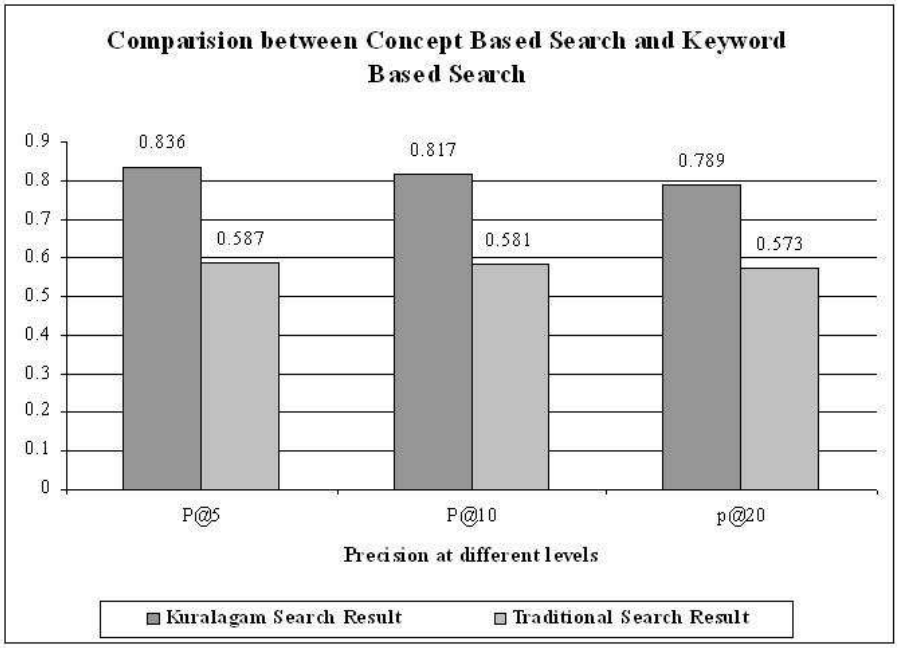

Fig 3: Average Precision Comparison

5. Conclusion and Future Work

This paper describes Kuralagam, a framework for concept relation based Thirukkural search in Tamil as well as in English using CoReX Techniques. Kuralagam unlike traditional keyword based Thirukkural searches retrieves Thirukkurals that are conceptually relevant to the Query. When compared to traditional search techniques, our conceptual search methodology has higher precision. In future enhancement can be made to increase the precision and recall score of the conceptual Thirukkural search.

Reference

1. Subalalitha, T V Geetha, Ranjani Parthasarathy and Madhan Karky Vairamuthu. CoReX: A Concept Based Semantic Indexing Technique. In SWM-08. 2008. India.

2. Foundation, U., the Universal Networking Language (UNL) Specifications Version 3 3ed. December 2004: UNL Computer Society, 2004. 8(5).Center UNDL Foundation

3. Anandan, R. Parthasarathi, and Geetha, Morphological Analyser for Tamil. ICON 2002, 2002.

4. T.Dhanabalan, K.Saravanan, and T.V.Geetha. 2002. Tamil to UNL Enconverter, ICUKL, Goa, India.

5. Andrew, T. and S. Falk. User performance versus precision measures for simple search tasks. In 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval 2006. Seattle, Washington, USA.