Using Human Plausible Reasoning as a Framework for Multilingual Information Filtering

Asma Damankesh, Jaspreet Singh, Fatima Jahedpari, Khaled Shaalan, Farhad Oroumchian

Abstract

In this paper the application of the theory of Human Plausible Reasoning (HPR) has been investigated in the domain of filtering and cross language information retrieval. The theory of Human Plausible Reasoning first has been introduced by Collins and Michalski on early 1990s; it has been applied to IR since 1995. This work is an extension to those experiments which focuses on building a framework for cross language information retrieval. The system built in these experiments utilizes plausible inferences to infer new, unknown knowledge from existing knowledge to retrieve not only documents which are indexed by the query terms but also those which are plausibly relevant.

Keywords: Human Plausible Reasoning, Plausible Inference, Information Retrieval Information Filtering

ACM Categories and Subject Descriptors: H.3.3 Information Search and Retrieval, Information filtering, Retrieval models

Introduction

From 1950's when the first Information Retrieval system has been implemented to date, several theories and techniques have been introduced and implemented by researchers in IR fields such as different document and query representation (i.e. Vector Space model, probabilistic models, Language modeling), query expansion and weighing functions. In all these works, the effort is to simulate the real-life behavior of information seekers and information finders. For example a reference librarian, although (s)he is not an expert in a particular domain, but can infer what books or documents could be relevant to an information seeker using general and even superficial knowledge of the subject. In this work an attempt is made to simulate the reasoning aspect of a reference librarian by modifying the theory of Human Plausible Reasoning.

Human Plausible reasoning is a relatively new theory for answering questions which is proposed by Collins and Michalski in 1989. Collins and his collegues have spent 15 years investigating how people can draw conclusions in an uncertain and incomplete situation by using indirect implications. They have developed a descriptive theory of human plausible inferences that categorizes the plausible inferences in terms of a set of frequently recurring inference patterns and a set of transformations on those patterns [1]. A transformation is applied on an inference pattern based on a relationship (i.e. generalization and specialization) to relate available knowledge to the query. Different experimental implementation of the theory such as adaptive filtering [2], XML retrieval [3] or expert finding [4] proves the flexibility and usefulness of HPR in the IR domain.

This research is about creating a framework for multilingual IR where all aspects of retrieval in this environment are represented as different inferences based on HPR. Our experiments so far has focused on the problem of retrieving relevant documents by the mean of plausible inferences as well as combining evidences of the relevance. In this system queries are processed and then represented as single words, phrases, logical terms and logical statements. Where a logical term represents a relation between two words and/or phrases and a logical statement represents a relationship between a logical term and one or more single word or/and phrase. Different inferences are applied on query terms to find documents indexed with these terms. In the next step these terms are transformed into new terms and their related documents are retrieved. The process of generating new terms from query terms or newly generated terms could be repeated several times. In this process, some documents will be retrieved through several inferences.

Each of these inferences are considered as an evidence of relevance and their weights are combined together in order to generate a single weight representing how much a document is relevant. In the context of the Information Filtering, the documents are user profiles and the queries are documents that are arriving one at a time. The problem we tried to address in our participation in CLEF this year was to measure the applicability of HPR multilingual filtering domain and to examine different methods for calculating the certainty and combining evidences of relevance. The attempt is made to build a framework which is independent of any specific language and can infer new knowledge which could be in a different language by utilizing relationships and general inferences.

This paper is structured as follow: in the first few sections briefly the theory of Human Plausible Reasoning (HPR) and the plausible inferences are described. Then the proposed system and inferences are explained. The experiments, findings and deficiencies of our implementation are explained next. The paper is concluded by providing guidelines for future research.

An Introduction to the Theory of Plausible Reasoning

For 15 years, Collin and his colleagues have been investigating the patterns used by people to reason under uncertainty and incomplete knowledge. They have concluded that these patterns could be categorized in terms of a set of frequently reoccurring inference patterns, and a set of transformation on those patterns. An inference applies a transformation on an inference pattern based on some relationship (i.e. generalization, specialization, similarity, dissimilarity) to relate available knowledge to the questions.

The theory assumes that a large part of human knowledge is represented in "dynamic hierarchies" that are always being modified, or expanded. [1] Concepts are represented by nodes, and are connected to each other by some relationships. Each node can belong to one or more hierarchy and in each hierarchy it'd viewed from different prospective. A node can be a clause (Lionas in Figure 1.b), an individual (Lionas in Figure 1.a) or a manifestation of an individual (Lionas in rainy season).

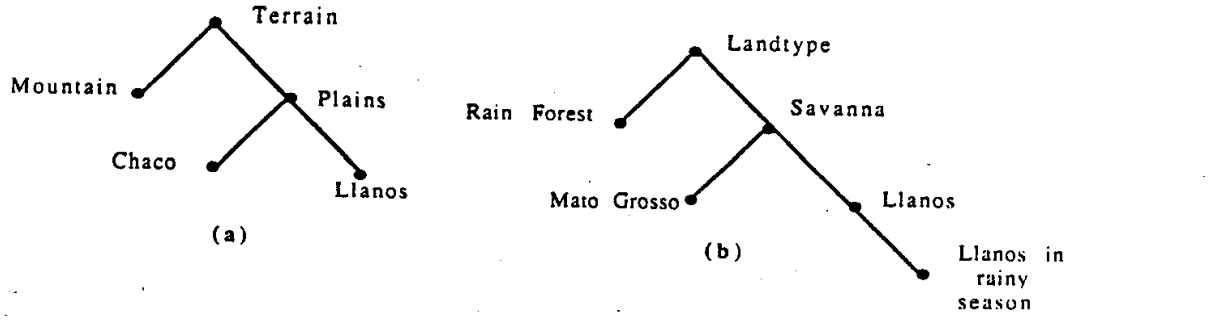

Figure 1

The primitives of the theory consist of basic expressions, operators and certainty parameters. In the formal notation of the theory, a statement like "Baghdad is the Capital of Iraq" is written as $capital(Iraq) = \{Baghdad\}, \gamma = 1.0$ where *capital* is descriptor, *Iraq* is an argument, *Baghdad* is a referent and $\gamma = 1.0$ is the certainty parameters that indicates we are 100 % sure that this fact is correct. The pair argument and descriptor is called *logical term. Logical statements* are terms associated with one or more referents. Descriptor, argument and referent could be any node in the hierarchy. In addition to the simple statements, dependencies can form logical expressions too. Elements of expression in the core theory have been summarized in Figure2. The theory has many parameters for handling uncertainty but it does not explain how these parameters could be calculated and this is left for implementations and adaptations. The definition of the most important parameters is given in Figure 3. The theory provides a rich set of inference transforms that could be applied on one statement to infer new knowledge from the available ones. Interested reader are referred to references [1] and [5] [6][7]

Baghdad is the capital of Iraq

referent r1, r2, {r2 ...}	e.g., Baghdad
argument a1, a2, F(a1)	e.g., Iraq
descritor d1, d2	e.g., Capital
term d1(a1), d2(a2)	e.g., capital(Iraq)
statement $d1(a1) = \{r1\} : \gamma, \varphi$	e.g., capital(Iraq) = baghdad :1,0.02

dependencies between terms

$d1(a1) \leftrightarrow d2(a1) : \alpha, \beta, \gamma$ e.g., latiude(place) ↔ average_temp(place) : moderate, moderate, certain)

(translation : i am certain that latitude constrains average temperature with moderate reliability and that average temperature constrains the latitude with moderate reliability.)

implicatios between statements :

$d1(a1) = r1 \Leftrightarrow d2(a1) = r2 : \alpha, \beta, \gamma$ e.g. grain(place) = {rice...} ⇔ rainfal(place) = heavy : high, low, certain

(translation : i am certain that if a place produces rice, it implies the place has heavy rainfals high reliability. but that if a place has heavy rainfall it only implies the place produces rice with loa reliability.)

Figure 2. Elemnts of expression in The Core Plausible Reasoning Theory

1. Certainty (γ) :	The degree of certainty or belief that an expression is true.
2. Frequency (φ) :	Frequency of the referent in the domain of the descriptor (e.g. a large percentage of birds fly.)
3. Typicality (τ) :	Degree of typicality of subset withing a set. (e.g. robin is a typical bird and ostrich is not a typical bird)
4. Dominance(δ) :	Dominance of a subset in a set(e.g. chickens are not a large percentage of birds but a large percentage of barnyard fowl.)
5. Similarity (σ) :	Degree of similarity of one set to another set.
6. Conditiona l Likelihood	(α) : conditiona l likelihood that the right − hand side of a dependency or implicatio n has a particular value (referent) given that the left − handside has a particular value.
7. Conditiona l Likelihood	(β) : conditiona l likelihood that the rleft − hand side of a dependency or implicatio n has a particular value (referent) given that the rightt − hand side has a particular value.
8. Multiplici ty of the referent(μr) :	*e*.g., many minerals are products by a country like Venezuela) .
9. Multiplici ty of the referent(μa) :	e.g., many countries produce a mineral like oil).
10. Acceptabil ity (A) :	users feedback

Figure 3 Certainty Parameters

Proposed System

Like any other logical system our system has four main elements which are document representation, query representation, domain knowledge base and a set of inference rules. Here, if the partial description of document can infer the query, we claim that the document is relevant to the query. Briefly, documents are partially described by concepts, logical terms and statements, and the knowledge base in created to hold relationships between concepts in the domain. Inference rules are continuously applied on the query term to expand it to infer other related concepts, logical terms and statement until a plausibly related document or documents are located.

Document Representation

In this model, documents are represented partially by a finite set of concepts, phrases, logical terms and statement that can be directly extracted from the document body or title. The reason why this representation is called partial document representation is that more terms could be inferred from existing ones that also represent the content of the documents. Every document identifies its concepts, logical terms and statements by the DOC relation. Three examples below indicate that the *doc#1* is about *the* concept *solider*, phrase *US_troop*, and the logical term *capital (Iraq)*

1. $DOC(solider) = \{doc\,\#1\}$
2. $DOC(US\_troop) = \{doc\,\#1\}$
3. $DOC(capital(Iraq)) = \{doc\,\#1\}$

Query Representation

Each query is processed similar to documents and is partially represented by its concepts, phrases and logical terms or statement. Each of these concepts, phrases, logical terms and statements are then form the argument of a logical term where the descriptor is the keyword DOC and the referent is unknown. So a query can be represented as a set of incomplete logical statements which have the form $DOC(partial-desciption) = \{?\}$

Therefore the retrieval process can be viewed as the process of finding referents and completing these incomplete sentences. Since a document can be retrieved in response to several terms in the query or through several inferences, therefore the final task is to combine the weights assigned to each document from each inference or term and create a sorted list of the retrieved documents for each query.

Document Retrieval by Plausible Reasoning

Like any other information retrieval system, in this system the first step is to find a direct match between the query representation and a document's partial description. That is, to locate the document or documents which are indexed by the query terms. Since the query is always represented as an incomplete statement, the aim of this direct approach is to complete the statement by finding the referents (documents indexed by the term). This direct approach is applied on each and every concept, phrase and logical term or statements which could be inferred from the query terms by applying the inference rule depicted in Figure 4.

$DOC(subject1) = \{?\}$

$DOC(subject1) = \{doc\#\}$: γ_1, A

– –

$DOC(subject1) = doc\#$: $\gamma = F_1(\gamma_1, A)$

subject can be a concept (e.g. Iraq), a phrase (US_troop) , a logical term (capital(Iraq)) or a statement (capital(Iraq) = Baghdad)

Figure 4 Finding references by completing incomplete query statement, Direct Approach

Another case is where the document is indexed by a concept, phrase, logical term or statement which is more specific or general case of the query term. The theory of plausible reasoning provide us with a rich set of transformations which could be applied on a concept, phrase, in the descriptor, argument or referent of a logical term or statement to convert the available statement to another which could be the index term of one or more documents. In this application of the theory only the GEN and SPEC inference transforms are used to move up and down the hierarchy. These inference rules are used only to infer new concepts and then the direct approach is applied on the new concept to retrieve the relevant document(s). Figure 5 illustrates the specialization (SPEC-) based argument transform by an example. As an example let's consider the query:

$$DOC(restaurant(iraq\_city)) = \{?\}$$

which indicates that there is an interest in documents about cities in Iraq which have restaurants. In the knowledge base we have the fact that the Baghdad is a city in Iraq. Therefore a new query term *restaurant (Baghdad)* will be added to the query representation as a new query term. Once the direct approach is applied on the new term *c* document *doc#3* is retrieved as a related document to the query.

$DOC(d(a)) = \{?\}$		$DOC(restaurant(iraq\_city)) = \{?\}$	
$a'SPEC\, a$	: δ_1, A_1	*Baghdad SPEC Iraq _ city*	: 1.0,1.0
– –		– –	
$d(a')$	: $\gamma_1 = F_1(\delta_1, A_1)$	$restaurant(Baghdad)$	: $\gamma = 1.0$
apply direct approach :			
$DOC(d(a')) = \{?\}$	: γ_1	$DOC(restaurant(Baghdad)) = \{?\}$	: $\gamma = 1.0$
$DOC(d(a')) = \{doc\#\}$	: δ_2, A_2	$DOC(restaurant(Baghdad)) = doc\#3$	: 0.6,1.0
– –		– –	
$DOC(d(a)) = doc\#$	: $\gamma = F_2(\gamma_1, \delta_2, A_2)$	$DOC(restaurant(iraq\_city)) = doc\#3$	$\gamma = 0.88$

Figure 5 Finding references by SPEC-based Argument Transform

The strength of our belief on the relevance of doc#3 to the query both depends on our belief on the suitability of the *restaurant (Baghdad)* as a representation for the query and how well that term is a representative of the content of the doc#3. Interestingly enough it would not make a difference if for example instead of "*Baghdad" we had "بغداد"* in our knowledge base. That is why we believe this approach is a general framework that can support multilingual retrieval.

A different case is when the document is indexed by a concept which is the referent of a query term. Or when the document is indexed by a logical term whose referent is a query term. Both cases are illustrated in figure 6 and figure 7 with examples.

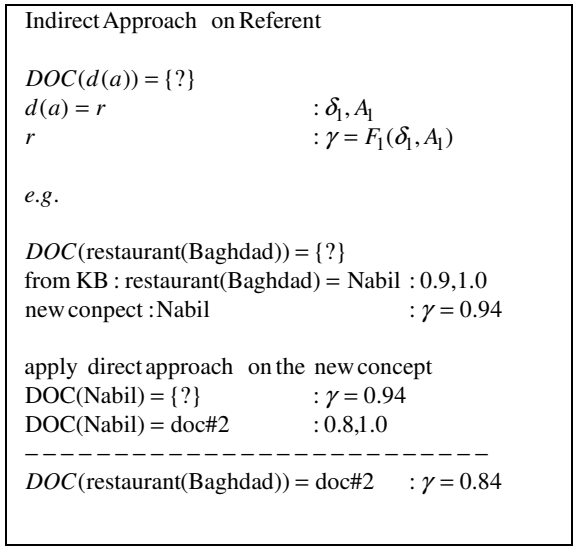

Figure 6 indirect approach on referent

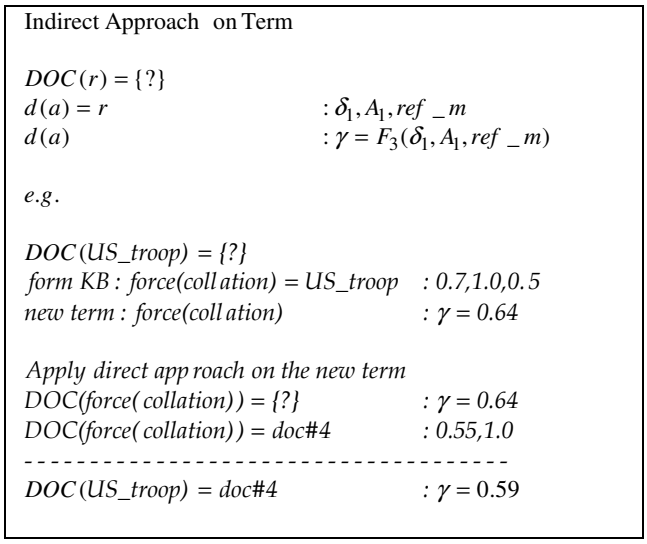

Figure 7 indirect approach on term

In each one of the above cases, first an inference has been applied to generate a new term and then the direct inference is used to retrieve the relevant document. The calculation of certainty parameters is discussed below.

Experiments

For the experiments on information filtering, first the collection was processed and its single words, phrases and logical terms and logical statements were extracted. The fact that which term came from which document was ignored at this stage. In the second step all the phrases were processed and some logical terms were also generated in this way. For example if we had a phrase such as abc. The following logical terms were generated c(ab) and bc(a).

Table 1 summarizes the number of tokens and relationships in the knowledge base.

Table 1- Number of tokens and relationships in the knowledge base

Token or Relationship Type	Count
Single words	143,512
Phrases	1,334,515
Logical Terms	78,964
Logical Statements	1,778,641
Kind of relationships	1,334,513

After creating the knowledge base, the profiles were processed and indexed as documents. Then each document was read, processed and match against the profiles using plausible inferences. The reasoning was limited to only two levels of depth because of speed limitations.

Unfortunately, we were not able to process all the documents and only processed the first xxx documents. This was due to the fact that this was our first major implementation in Python, and we learned the implementation issues the

hard way! We were not able to run the infile client although we received help from Romaric Besançon and INFILE team but still we had to run the file version of the client. Because we ran out of time, we were not able to work on thresholds and refining the output of the system, so our results were generally poor. But we hope we will be able to do much better next year.

Conclusion

In this work an attempt has been made to adapt the Collins and Mechaslki's theory of Human Plausible Reasoning as a multilingual framework for information retrieval and information filtering. In these experiments we were able to build the relation extractor to build a knowledge base, document processor and query processor based on plausible inferences. However, due to time limit and implementation issues with Python we were not able to add Arabic to our knowledge base. Also, we were not able to demonstrate a reasonable performance by the time of the CLEF deadlines. However, we were able to show how this approach could be used to handle multiple languages.

There are many potential improvements to the current system. First is enriching the knowledge base by implementing better NLP techniques with the ability to produce more accurate and reliable set of terms and statements. We need to improve our Arabic text processing to form more logical terms and statements. We need to experiment with different methods of calculating certainty of inferences and combining evidences because currently we implemented the simplest methods. Other suggestions are applying context, using heuristic and machine learning strategies. HPR allows defining context for each one of the relationships in the knowledge base and uses them in inferences. This improves the quality of inferences which is really needed in an information filtering situation.

References

[1] ALLEN COLLINS, R. MICHALSKl, 1989, "The Logic Of Plausible Reasoning A Core Theory", Cognitive Science, Vol. 13, pp. 1-49.

[2] F. Oroumchian, B. Arabi, E. Ashouri, 2002, "Using Plausible Inferences and Dempster-Shafer Theory of Evidence for Adaptive Information Filtering.", 4th International Conference on Recent Advances in Soft Computing, Nottingham, United Kingdom

[3] M. Karimzadehgan, J. Habibi, F. Oroumchian, 2005, "Logic Based XML Information Retrieval for Determining the Best Element to Retrieve", Computer since Springer, pp88-99

[4] Maryam Karimzadehgan, Geneva G. Belford, and Farhad Oroumchian, Expert Finding by Means of Plausible Inferences, International Conference on Information and Knowledge Engineering (IKE'08),Las Vegas, USA, July 14-17, 2008.

[5] ALLEN COLLINS, MARK H. BURSTEJN, 1988, "Modeling A Theory Of Human Plausible Reasoning", Artificial Intelligence III.

[6] Maryam Karimzadegan1, Jafar Habibi, Farhad Oroumchian, "XML Document Retrieval by means of Plausible Inferences ", in Advances in XML Information Retrieval, Third Workshop of the INitiative for the Evaluation of XML Retrieval INEX 2004, Schloss Dagstuhl, 6-8 December 2004, Lecture Notes on Computer Systems, , Editors N. Fuhr, M. Lalmas, S. Malik and Z. Szlávik, Springer Verlog LNCS 3493, ISBN: 3-540-26166-4, 2005.

[7] Ehsan Darrudi, Masud Rahgozar, Farhad Oroumchian, "Human Plausible Reasoning for Question Answering Systems", International Conference on Advances in Intelligent Systems - Theory and Applications in cooperation with IEEE Computer Society - AISTA'2004, Luxembourg, 15-18 November 2004.