Information Retrieval Project: Introduction to Fuzzy Logic and Fuzzy Information Retrieval

Bruno Bonaiuto Bolivar

Course of AA 2022-2023 - DSSC

Abstract

This work introduce the idea of a fuzzy logic system, the fuzzy control and how the fuzzy logic works. Also since the fuzzy set theory is possible to implement in many areas, this work introduce to the use of fuzzy model for information retrieval systems this approach to capture the relationships between words and query language.

Introduction

fuzzy logic derives from the fact that most of human reasoning are approximate in nature. The development of fuzzy logic was motivated by the need for a conceptual framework which can address the issues of imprecision. The benefits of this technique are, it is more flexibility and generality in the formulation and solution problems. The fuzzy logic is suitable for linear and non linear problems.

An information retrieval system stores and indexes documents in that way when the user requires some information in a query system, it retrieves the related documents according to some score to each one. The higher the score and the greater is the importance of the document. Moreover, documents are retrieve when the contain the index terms specified in the queries. To deal with the the **vagueness** of humans, The fuzzy set can be used to modify the knowledge in the bases.

The Expectation is in the indexed terms, improving the quality of retrieved documents in order to get the most relevant and more semantically related to the initial query. There are some ways in performing this form of search, as Full-text search one of the most popular for retrieving documents providing keywords to search for. But this comes with some problems as dealing with thousands of documents.

Apache Lucene has been also used for retrieve relevant documents. Lucene is a library used for implementing full-text search, it performs a fuzzy search in two steps. First, search for tokens stored in the database with similarity to the query tokens by computing the edit distance. Second, it uses the similar tokens it finds as new query tokens to retrieve relevant documents.

In this paper, we propose and introduction to fuzzy logic and fuzzy IR.

1 Fuzzy set

A fuzzy set is a class of objects with a continuum of grades of membership (characteristic) function which assigns to each object a grade of membership ranging between zero and one. it provide a natural way of dealing with problems.

A fuzzy set is allow any set to have freedom in the degree of membership, this is called the membership function [0,1], The notions of inclusion, union, intersection, complement, relation, convexity, and others, are extended as sets.

Considering the statement of "Carl is 1759.5999 mm high". It might seen a very accurate statement but instead it depends on the field when it is presented. In a environment of Oriental people Carl might be seem as a tall person, but instead changing environment with NBA player Carl could be seem as a short person. To formalize these aspects Zadeh introduced the concept of **Linguistic Variable**

A linguistic Variable has a name, a definition domain, a set of values and an interpretation. This means that the name of a variable can be chosen freely. The definition of Domain depends with the universe where it will be used. Linguistic terms are called fuzzy sets.

Fuzzy logic is basically a logic approach which allows intermediate truth values to be defined between conventional evaluations of true and false. Some notions as rather hard or pretty cool can be formulated mathematically and processed by computers.

Fuzzy logic derives from the fuzzy set theorem differs from the traditional logical systems truth or false, where many degree of membership are allow, in this way fuzzy logic is able of inherently imprecise concepts. Fuzzy logic is, which is the logic on which fuzzy control is based. Essentially it provides an approximation of the inexact nature of the real world.

Fuzzy logic control, provides a nonlinear algorithms, characterized by a series of linguistic statement, into the controller.

1.1 The membership function

A graph that defines how each point in the input is mapped to a membership with a value between 0 and 1. The Input space is referred as universal set (u), containing all the possible elements in each particular application.

Deciding which of the membership functions to use is important in the design of fuzzy logic controller. The most popular membership functions are: Triangular, Gaussian, Trapezoidal

1.2 Fuzzy logic Control

The structure of the fuzzy logic control consist on the following parts:

- Fuzzification,
- Knowledge base (rule base),
- Inference engine,
- Defuzzification

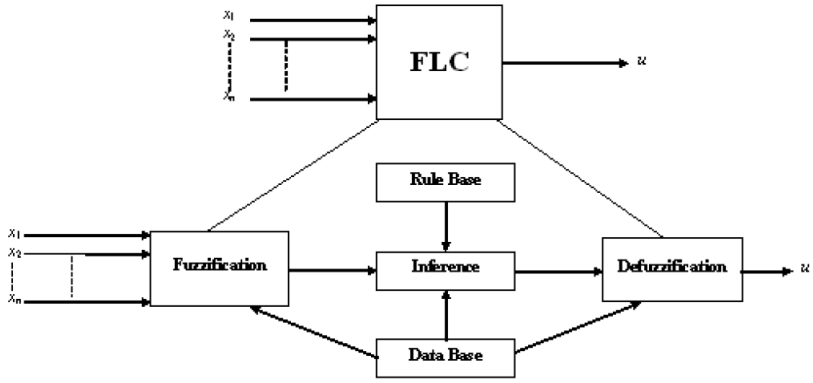

Figure 1: Fuzzy Logic Control (Architecture)

The fuzzification module converts the crips values of the control input into fuzzy values. A fuzzy variable has value which are defined by linguistic variables (fuzzy sets) such as low, medium, high, big, slow, etc. Where each one is defined by a gradually varying membership function.

The Rule base, is the control strategy of the system, which is obtain by expert knowledge or heuristics, it is composed by conditional statements expressed as a set If-Then rules.

The inference engine, determines the matching degree of the current fuzzy input with respect to each rule and decides which rules are to be fired according to the input field

The defuzzification, a mathematical procedure of converting fuzzy values into crips values, There are several defuzzification methods available and the best-suited one is used with a specific expert system to reduce the error.

Therefor a fuzzy model provides the behaviour of a system to help the user in the understanding of the system, and if the model use formalism of fuzzy logic its called a fuzzy model.

The composition operation is the method by which such a control output can be generated using the rule base. There are several composition methods, max-min or sup-min and max-dot and others.

1.3 Advantages and Disadvantages

In this section are listed some advantages and disadvantages for the fuzzy based systems.

1.3.1 Advantages

- High precision on outcomes.

- The fuzzy logic is similar to human reasoning (understandable).
- Usages simple arrive with simple mathematical models,for solving real world linear or non-linear problems.
- Efficient and decision control.

1.3.2 Disadvantages

- Low speed and long running time required.
- For precise results the system needs more data, and it also needs to increase the rules of reasoning.
-Not usages the feedback of system.

1.4 Applications

Machine learning and pattern recognition: in problems such as classification, clustering, optimization and others, since the data needs to be evaluated, captured for patterns, and extracted from raw data. Therefor fuzzy logic is suitable for machine learning and pattern recognition, in areas as image recognition, text classification, and others.

Fuzzy logic can be helpful for improving existing models. For **text mining**, in a number of places are available now days such as opinion mining, feedback management and others. In **Object Recognition**, finding the objects in theses frames can be suitable and effective for fuzzy logic.

Also for image classification, image retrieval, and optimization.

2 Boolean information retrieval Model

Now days, most of the information retrieval systems are based on the Boolean logic model, which assumes that the query provide by the user describes properly the user needs, With no scope for vagueness or fuzziness. However this is inappropriate due to the fact that the user might include fuzziness in the queries. The reason behind is that the user might not know enough about the subject or the query might not describe the information properly.

Other issues such as performance, scalability are other common information retrieval issues. These algorithms define how relevant is a document to a user query by using some define functions between the query and the document in the index.

Therefor, managing volumes of information from the web, and trying to retrieve most relevant documents to a query, may be difficult and can return thousands of documents, being many of these documents loosely related to the original retrieval criteria .

Since fuzzy set theory can be used to describe imprecise information, is more practical to apply the fuzzy set theory to information retrieval systems. An information retrieval system that have a good query management with the

ability to give weight to documents that are more relevant to the user's query, and present those results first.

3 Fuzzy Information Retrieval

The fuzzy information retrieval systems uses tools defined in fuzzy logic and fuzzy relations in order to get the best results to a user query. In contrast to Boolean systems, Fuzzy systems are better with dealing with data that may display a degree of membership. In fuzzy systems, objects describe with properties are assigned relational membership values to show: relevancy from properties and vice versa. This provide values different than probabilistic systems. Whereas in fuzzy systems, a membership value is used to dertermine a weighted relational mapping.

The fuzzy information retrieval model is fuzzy generalizations of the Boolean model. The fuzzy information retrieval assumes that a set of fuzzy documents is associated with each word in the query language. That is to say, each word in the query language defines a fuzzy set, and the elements in the sets are retrieved documents.

Correspondingly, each document in the set has a degree of membership to correspond to each word in the query language. As a retrieval result, the fuzzy set reflects how well each document matches the query.

Indexing is the preliminary operation in the creation of the documents' representation. In order to define an index, we should ensure that the index can present textual information. An indexing function, as the membership function of the fuzzy set, and is used to calculate the correlation between words.

In order to understand better the whole idea, let's define some basic concepts:

- Document indexing The indexing process is a procedure to assign a set of descriptive words (index terms) or phrases to synthesize what about is a document. This process can be performed manually by an expert in the field or automatically.

- Queries The procedure of requesting information needed, is called a query, normally the users formulate queries in natural language, describing those needs. If a system is expected to respond, it will need to analyse those natural language statements.

With that said, The basic idea is to expand the set of index terms in the query, with the related terms, such that additional relevant documents can be retrieved.

Since it's needed the notion of proximity among index terms, the following function achieve the goal to presents textual information properly.

3.1 The membership function F

$$\{(t, F(d,t)) \mid t \in T\} \, for \; d \in D, \tag{1}$$

Where $F(d,t)$ is the membership function that computes the degree of correlation between each index term and each retrieval document d. Therefor, T and D are the collection of each item in the query language and retrieval documents.

Commonly it is used as a definition of the membership function F based on term frequency (tf) and inverse document frequency ($id\,f$) in the following form:

$$F(d,t) = tf \ \times \ idf, \tag{2}$$

where

$$tf = \frac{the\ number\ of\ occurrences\ of\ query\ term\ t\ in\ d}{The\ number\ of\ all\ words\ in\ d}, \tag{3}$$

$$idf = log(\frac{the\ total\ number\ of\ documents\ in\ the\ retrieved\ set}{the\ number\ of\ documents\ indexed\ by\ query\ term\ t\ +\ 1}). \tag{4}$$

In other words, tf is the frequency of index term t in document d, and idf is a measure of the importance of the index term t. From Equation (3) and (4), we can appreciate that the value of $F(d,t)$ will be a little high if a term is high frequency for a given document and low frequency for the whole of the retrieved documents.

3.2 Techniques

the fuzzy string searching technique can be formulated as: "find in the text of size n all the words that match with a given word, taking into account k possible differences(errors)."

In most cases a metric is understood as a more general concept. This concept can also be called distance.

3.2.1 Levenshtein distance

Also known as "edit distance", is the most used metric, the algorithms of its computation can be found at every turn. And the entire process can be represented in a matrix.

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Figure 2: Levenshtein distance formula

3.2.2 Prefix Distance

Used to calculate the distance between the prefix pattern and a string, by taking the smallest of the distance from the prefix pattern to all the prefixes of the string.

3.2.3 Damerau - Levenshtein distance

This technique is a variation of the Levenshtein distance, by adding one more rule. "transposition of two adjacent letters are also counted as one operation, along with insertions, deletions, and substitutions". In practice this metric provide the best results.

$$d_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 & \text{if } i,j>1 \text{ and } a_i = b_{j-1} \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} & \text{and } a_{i-1} = b_j \\ d_{a,b}(i-2,j-2)+1 \end{cases} \\ \min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 & \text{otherwise.} \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \end{cases}$$

Figure 3: Damerau-Levenshtein distance implementation process

3.3 Approach

As noted, the fuzzy model accepts documents statistic and produce one relevance value representing the degree of document to be considered query. As show in figure 1. The fuzzy system consist is applied to this form introduced by L. Zadeh in 1965, In the fuzzy logic will be any value in the interval that start by zero and ends by at one. The fuzzy way of representing the relevance is given by a degree , we can have high, medium and low relevance fuzzy. In order to reach a higher precision it could be added some extra degrees as very high and very low.

In the membership function calculated by the function F shown in equation 1. We can add more variables as document frequency ratio *df* or the overlap. More fuzzy value can be added as high relevance, medium and low relevance as mentioned above, and this stage of deciding the fuzzy variables, fuzzy values, the fuzzy membership functions and getting the membership values is called **Fuzzification** in the context of the fuzzy information Retrieval.

The next stage is called the **FIS** stage, where depending on the fraction of corpus documents that has the term t appearing in it. These fuzzy values

should be related to the documents using intuitive rules or based on experience. As an example assuming three rules as a sample of fuzzy rules, providing two relevance values (high and low), will be as follows:

- if ((high-tf) And (low-df)) then high-relevance . . . rule 1.
- if ((medium-tf) And (low-df)) then high-relevance . . . rule 2.
- if ((low-tf) And (high-df)) then low-relevance . . . rule 3.

Subsequently, In fuzzy logic the operators are redefined according as minimum, maximum, and complement. They are called the *Zadeh operators*. According to the definition of Zadeh operators, we take the minimum of any tow Added values then we take the maximum of the relevance values that belong to the same fuzzy value. We should end up with three relevance values as: high, medium and low relevance. It will be as the following example:

Assuming that (high-tf = 0.7, medium-tf =0.4, low-tf =0.1, high-df = 0.3, low-df=0.6).

- high relevance = *minimum* (0.7, 0.6) = 0.6 . . . .this is after applying rule 1.
- high relevance = *minimum* (0.4, 0.6) = 0.4 . . . .this is after applying rule 2.
- low relevance = *minimum* (0.1, 0.3) = 0.1 . . . .this is after applying rule 3.

Now we have two fuzzy values for the considered document representing its degree of high and low relevance, the document is 0.1 low relevant and 0.6 high relevant to the query.

The last stage called the **Defuzzification** which is moving from fuzzy values into one single value that represents the relevance of the documents. The defuzzification can be done by applying the centroid function or the weighted average method. As an example, the weighted average is calculated by combining the deduce values for high and low relevance and doing some sums and multiplications operations, it will result in a single value as follows:

$$Relevance - value = \frac{1 \times \mathbf{0.6} + 0.1 \times 0.1}{1 + \mathbf{0.1}} = 0.554 \tag{5}$$

As it's shown in equation 5. only two documents fuzzy variables were considered, the term frequency tf and the document frequency df and we ended up with 0.554 as the relevance value.

3.4 Performance Evaluation

Recall and precision measures are frequently used to evaluate the effectiveness of the information retrieval systems

3.4.1 Precision

It is a fraction of documents that are relevant among the entire retrieved documents. Basically it provides accuracy of result.

$$Precision = \frac{Set\,of\,relevant\,documents\,retrieved}{Set\,of\,documents\,retrieved}. \tag{6}$$

3.4.2 Recall

A fraction of the documents that is retrieved and relevant among all relevant documents. Practically, it gives coverage of result.

$$Recall = \frac{Set\ of\ relevant\ documents\ retrieved}{Set\ of\ all\ relevant\ documents}. \tag{7}$$

4 Conclusions

The fuzzy logic is an essential part of technology, that offers services to solve real world complexities. The solving technique is transparent and flexible to scale a minimize the requirements according to the needs. Subsequently, we discussed details of an IR system with fuzzy logic. This is a new technique having advantages over the information retrieval system as it can handle vague and imprecise queries of user very well. The fuzzy notions describe situations known through imprecise, uncertain, and vague information in a way that neither replaces nor is replaced but that, instead, complements the views produced by other approaches.

References

[1] B. Karn, "Information retrieval system using fuzzy set theory–the basic concept," *Assistant Professor, Department Of MIS (Management Information Systems), Birla Institute Of Technology, Mesra, Ranchi*, 1998.

[2] D. Qiu, H. Jiang, and S. Chen, "Fuzzy information retrieval based on continuous bag-of-words model," *Symmetry*, vol. 12, no. 2, 2020.

[3] C. Moraga, "Introduction to fuzzy logic," *Facta universitatis - series: Electronics and Energetics*, vol. 18, pp. 319–328, 09 2005.

[4] V. Kushwah and A. Bajpai, "Importance of fuzzy logic and application areas in engineering research," 03 2019.

[5] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[6] A. abdel ghani and A. Tahour, *Application of Fuzzy Logic in Control of Electrical Machines*. 03 2012.

[7] N. Thakur, P. Singh, S. Dhawan, and S. Agarwal, "Implementation of an efficient fuzzy logic based information retrieval system," in *Bharati Vidyapeeth's College of Engineering/Computer Science*, 2015.

[8] L. N. Al Aziz, *Enhancement of information retrieval ranking using fuzzy logic*. PhD thesis, The British University in Dubai (BUiD), 2011.