Cell Assemblies for Query Expansion

Isabel Volpe and Viviane Moreira

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
{icvolpe, viviane}@inf.ufrgs.br

Abstract

This paper applies the Cell Assemblies (CAs) model to Information Retrieval. CAs are reverberating circuits of neurons that can persist long beyond the initial stimulus has ceased. CAs are learned through Hebbian learning rules and have been used to simulate the formation and the usage of human concepts. We adapted the CAs model to learn relationships between the terms in a document collection. The method will be validated by means of experiments on standard IR test collections.

Key-words: Query Expansion, Information Retrieval, Cell Assemblies.
1. Introduction

Information Retrieval (IR) deals with representation, storage, organization, and access to information items [Baeza-Yates and Ribeiro-Neto 1999]. The IR process starts with a user who has an information need. The user typically translates this need into a query composed of a set of keywords. These keywords are then submitted to an IR System (IRS) which retrieves the items (textual documents, web pages, images, videos, etc.) that are likely to satisfy the user’s information need.

Query Expansion (QE) is one of the most widely used methods for solving the synonymy problem (i.e. the fact that the same concept may be expressed using different words). The basic idea is to augment the original query with synonyms and related terms in order to increase the number of relevant documents retrieved.

The main contribution of this work is to propose a new method for QE based on the Cell Assembly Model (CAs). The CAs model proposed by Hebb [Hebb 1949] implies that groups of active neurons in the brain are responsible for the storage of knowledge of human beings. There is substantial agreement that they form the basis of the human concepts and solve a wide variety of other problems [Huyck and Orengo 2005, Sakurai 1998]. Several experiments simulated a range of behaviors by which Cell Assemblies are responsible [Huyck 2004, Ivancich, Huyck, et al. 1999].

In a preliminary study, [Huyck and Orengo 2005] have shown that CAs can be used to perform categorization and IR, proving its potential for improving performance when used in an IRS. However, this study was only exploratory and a series of questions about the use of CAs in IR remain unanswered. In this dissertation we propose using CAs for QE.

This paper is organized as follows: Section 2 summarizes related work on the topics QE and the use of Neural Networks for IR; Section 3 introduces the CAs model; Section 4 presents our proposal of using CAs for QE; finally, Section 5 summarizes our work and presents our final considerations and future work.

2. Related Work

This section briefly revises related work on Query Expansion and Neural Networks for IR.

2.1. Query Expansion

An IRS requires a precise and comprehensive query in order to perform the search and ranking of documents so that only relevant documents are presented to the user. However, the specification of the query is limited by the user's vocabulary and knowledge of the search domain. Query Expansion (QE) aims to retrieve not only documents containing the query terms but also documents that have related terms [Manning, Raghavan, et al. 2008]. It adds terms that have a semantic relationship with those in the original query. The purpose is to enable the retrieval of documents, even if they do not have words that were present in the original query. What differentiates the types of query expansion is the method by which these additional terms are chosen. The central question is how to generate expanded queries. Two approaches can be taken:

(i) **Local methods** use information from the set of documents retrieved by the original query to choose expansion terms. These methods usually require user intervention to mark a few documents as relevant. This process, known as Relevance Feedback (RF), typically achieves significant retrieval improvements. Experiments by Salton & Buckley [Salton and Buckley 1997] on small data collections report improvements from 47% to 160%. However, since RF requires effort from the user, it is not commonly used in working systems [Spink 1994].

(ii) **Global methods** expand the query without taking the results retrieved by the original query into consideration. This is usually achieved with the aid of a thesaurus or WordNet.
A thesaurus is a controlled dictionary in a given domain of knowledge and is used for identifying synonymous expressions and linguistic entities that are semantically similar. The basic procedure is for each term in a query, automatically expand with synonyms and related words from the thesaurus. The main advantage of this method is not requiring user intervention. The main limitation is that building a thesaurus manually is very costly [Manning, Raghavan, et al. 2008, Zhang, Deng, et al. 2009]. The WordNet is organized in sets of synonyms (synsets) with the words of same meaning allowing searches for semantically related nodes. [Fellbaum 1998, Grootjen and Weide 2006, Parapar, Barreiro, et al. 2005] used WordNet as a source of additional words to complement the user's query. All three studies report that no significant improvement was obtained with the expansion.

2.2. Neural Network Models for IR

In this section we present some studies that use Artificial Neural Networks (ANN) as an alternative IR model. The main motivation for the use of ANNs in IR is that they perform well at pattern matching tasks, and this ability can potentially aid document retrieval.

ANNs are a computational model that simulates the processing of information in the human brain. They are composed of simple processing units, the neurons, which are connected via synapses which have a weight. These synaptic weights reflect the level of association between the two neurons they connect. The collection of artificial neurons forms a network [Scholtes 1991].

An ANN resembles the brain in two aspects: (i) synaptic weights are used to store knowledge; and (ii) knowledge is obtained via learning cycles [Roberson and Dankel II 2007]. Despite the two systems being very different in terms of biology, these features allow the ANN to faithfully reproduce many functions which are only found in the human brain [Haykin 2001].

One of the first attempts to use an ANN for IR was done by [Belew 1989] who proposed a three-layer network with authors, index terms and documents. His system employed relevance feedback from the users to generate a consensual representation of the meaning of keywords and documents shared by the group of users. Roberson and Dankel II [2007] used a Morphological NN (MNN), which differs from other ANNs by the way processing occurs in the nodes. Multiplication and addition are replaced by addition and maximum (or minimum), respectively. As a result MNN computation is nonlinear before thresholding. Retrieval experiments on the TIME test collection reported results which were significantly worse than the vector space model.

3. Cell Assemblies Model

CAs are groups of neurons that have a substantial synaptic strength. It is a network reverberation that can remain active even after the external stimulus has ceased [Huyck 1999]. The relationship between the terms can be learned by specifying the network topology based on a collection of texts. With the running of the network, documents can be retrieved from the network of relationships created.

ANNs are popular computational models that are said to be inspired by human neural functioning. However, natural systems work quite differently from most ANNs. The CANT [Huyck 1999] model is a CA model that has been used for IR [Huyck and Orenge 2005], learning hierarchical categories [Huyck 2007], and as a video game agent [Huyck and Byrne 2009], among other applications.

The processing is broken into discrete time steps. On each time step, any neuron that has an activation level greater than a given threshold will fire, and the activation levels for all of its post-synaptic neurons are updated. Neurons are connected by a small number of synapses. Neurons in the network are sparsely and randomly connected via synapses to 40
other neurons [Huyck and Orengo 2005]. In order to examine the firing behavior of the network their activity is plotted as a function of time. Figure 1 shows the network as a rectangular matrix. The neurons that fired are shown in blue.

![Figure 1. Example Network Activation](image)

The model was designed to have an operation similar to that of natural neural systems. The basic idea is derived from Hebb's [1949] idea of the Cell Assembly which is the neural equivalent of a concept. At any given run of the network, must lead to the activation of the proper CA.

The CANT model generates CAs from a variety of parameter settings. It has six key neural properties, the first three are fairly common in ANN models but the latter three are less common [Huyck 1999]:

1. **Connection Strength**: neurons have connections to other neurons. Connections may have positive or negative strength and the continuous activation is simulated by time steps.
2. **Activation**: when the neuron crosses a threshold, it sends activation down each of its axons. In this situation, we say the neuron **fires**.
3. **Activation Threshold**: a neuron fires if has enough activation to surpass the activation threshold.
4. **Connectivity**: the closer the neurons, the greater the probability of being connected. The model simulates the distance-bias connectivity between the neurons.
5. **Decay**: each step decreases the activation, but a new activation can lead to a net gain of activation. Decay is a constant that applies to active and inactive nodes.
6. **Fatigue**: the more time that the neuron is active the higher the threshold becomes. Fatigues will make the neuron less likely to remain active.

### 3.1. Neurons

Neurons receive activation from other neurons via synaptic connections. A neuron fires only if it achieves sufficient activation to exceed the activation threshold. When a neuron fires, it sends activation through its synapses to other neurons [Passmore and Huyck 2008] and loses all its activation. However, if a neuron does not fire, some of its activation leaks away. A neuron should not be able to fire repeatedly. Thus, the model also allows neuron fatigue. This is modeled by raising the activation threshold each time a neuron fires. When a neuron does not fire, its activation threshold decreases at each time step until it reaches a given base level [Passmore and Huyck 2008].

### 3.2. Hebbian Learning

The CA model uses a Hebbian Learning rule which states that if two neurons fire at the same time step, the strength of their synapses should be increased. The two neurons are called pre-synaptic and post-synaptic.
For IR purposes, CAs also uses a compensatory learning rule. This rule limits the total synaptic strength of a neuron. As a result, the neurons that represent rare terms will have increased its influence and neurons that represent very frequent terms have their influence reduced. This rule is similar to IDF (inverse document frequency) weighting widely used in IR [Huyck and Orengo 2005].

4. Cell Assemblies for Query Expansion
The key motivation for using CAs for QE is to explore its spreading activation to augment the original query with related terms. The computational model is based on mammalian neural processing. The neurons are connected via uni-directional synapses and learning occurs by synaptic modification. Unsupervised learning rules change the synapses and the changes are based solely on the properties of the pre and post-synaptic neurons.

In the CANT simulations discussed in this work, each term (after stemming and stopword removal) is represented by a neuron and, the documents are represented by sets of neurons. By representing a term using one neuron, the size of the network is markedly reduced enabling the encoding of several thousand words and learning of the relationships between many documents.

Each term that occurs in more than one document was assigned a neuron. Each neuron is connected to other neurons representing the words that it co-occurs with at least once across the document collection. The selection of these terms is made randomly.

The training phase is the first step of the process, in which documents are presented to the CA network. The neurons corresponding to words that are present in the documents receive external activation. As the documents are presented, the weights of the synapses between pre- and post-synaptic neurons are adjusted. The words are directly mapped to the neurons, thus, the network learns the relationships between words. At the end of this phase, the semantic relationships among terms are modeled. These semantic relations are derived from characteristics of the distribution of terms in the document collection. This method is based on the assumption that term co-occurrence statistics provide useful information about the semantic relationships between terms.

In the querying phase, the queries are presented to the CA network. The neurons that represent the terms in the query receive external stimulation. As a result they fire and send activation to other neurons through synaptic connections. This activation is distributed for a few cycles and the state of the network is saved. This process will have the effect of expanding the original query with correlated terms. The rationale is that by adding correlated terms, the IR system will retrieve more relevant documents.

The experiments intend to assess the feasibility of employing CAs for QE. For the baseline run, the original text from the topics will be submitted to the IR system as queries. For the QE run, the neurons in the CA network that correspond to the original query terms will be externally stimulated. The activation will spread for five cycles, then the status of the network will be saved. The terms corresponding to the neurons that were firing after five cycles of spreading activation will be added to the original query. The performance of the baseline run will be compared to the performance of the QE run. The retrieval measure used will be mean average precision. In order to calculate this measure, we need an IR test collection containing documents, queries, and relevance assessments. There are a few test collections available, in particular, we plan to use the LA Times collection part of the CLEF Test Suite available from ELDA\(^1\). In order to assess if the gain in retrieval quality brought by QE was significant, we will employ statistical tests.

\(^1\) http://www.elra.info/.
5. Conclusions

The CA network was designed to model the computationally important aspects of mammalian neural processing. CA networks are a model of human neural processing that is based on simple neurons that are connected by synapses that learn via Hebbian learning.

This work proposes the use of CAs for QE. Future work includes experiments to validate the proposal.

Acknowledgments

This work was partially supported by CNPq.

References

Scholtes, J.C., *Neural Nets and their relevance for Information Retrieval*, Department of Computational Linguistics, University of Amsterdam 1991.