UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ISABEL CRISTINA VOLPE

# *Cell Assemblies* para Expansão de Consultas

Dissertação apresentada como requisito parcial
para a obtenção do grau de
Mestre em Ciência da Computação

Prof[a]. Dra. Viviane Pereira Moreira
Orientadora

Porto Alegre, junho de 2011

**CIP – CATALOGAÇÃO NA PUBLICAÇÃO**

Volpe, Isabel Cristina

*Cell Assemblies* para Expansão de Consultas / Isabel Cristina Volpe. – Porto Alegre: PPGC da UFRGS, 2011.

51 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2011. Orientadora: Viviane Pereira Moreira.

1. Expansão de consultas. 2. Recuperação de informações. 3. Redes neurais. 4. Aprendizado *hebbiano*. I. Moreira, Viviane Pereira. II. Título.

*“A felicidade não depende do que nos falta, mas do bom uso que fazemos do que temos.”*

— THOMAS HARDY

# AGRADECIMENTOS

Agradeço a todos que contribuíram para o desenvolvimento desse trabalho. Em especial gostaria de agradecer:

A Professora Viviane pela orientação conduzida, pelos conselhos e pela atenção que me foram dados. Agradeço pela confiança e pela oportunidade de cursar o mestrado.

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte financeiro. Ao Instituto de Informática da UFRGS por toda a infraestrutura oferecida e, ao PPGC e seus funcionários.

Aos professores agradeço pelos ensinamentos transmitidos. Aos professores Ana Bazzan, Aline Villavicencio e Marcelo Pimenta agradeço pelo interesse que tiveram em cada etapa desta jornada.

A Professora Renata Galante pelos incentivos e motivação para ingressar no mestrado, e por sempre estar disposta a ajudar quando surge qualquer dúvida.

A toda a minha família e aos meus amigos que acompanharam, de perto ou de longe, a realização deste trabalho e que torceram para que o mesmo fosse concluído com êxito.

Aos meus colegas de laboratório que são ótimos amigos e que propiciaram um ambiente de integração e troca de experiências, que partilharam momentos de descontração e seriedade durante estes dois anos.

Por fim, agradeço ao Rodrigo pela ajuda, amor, paciência e companheirismo ao longo destes 17 anos de convivência. Agradeço por entender as minhas ausências e por ter me propiciado alcançar mais esta meta.

# SUMÁRIO

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AIR | Adaptative Information Retrieval |
| CA | Cell Assemblies |
| CLEF | Cross-Language Evaluation Forum |
| CANT | Connections, Associations and Network Technology |
| EC | Expansão de Consultas |
| FLIF | Fatiging Leaky Integrate and Fire |
| LA Times | Los Angeles Times |
| MAP | Mean Average Precision |
| RI | Recuperação de Informações |
| RN | Rede Neural |
| RNM | Redes Neurais Morfológicas |
| SRI | Sistemas de Recuperação de Informações |
| TF-IDF | Term Frequency - Inverse Document Frequency |
| TREC | Text Retrieval Conference |
| VSM | Vector Space Model |

# LISTA DE FIGURAS

# LISTA DE TABELAS

# RESUMO

Uma das principais tarefas de Recuperação de Informações é encontrar documentos que sejam relevantes a uma consulta. Esta tarefa é difícil porque, em muitos casos os termos de busca escolhidos pelo usuário são diferentes dos termos utilizados pelos autores dos documentos. Ao longo dos anos, várias abordagens foram propostas para lidar com este problema. Uma das técnicas mais utilizadas, com o objetivo de expandir o número de documentos relevantes recuperados é a Expansão de Consultas, que consiste em expandir a consulta com a adição de termos relacionados. Este trabalho propõe um método que utiliza o modelo de *Cell Assemblies* para a expansão da consulta. *Cell Assemblies* são grupos de neurônios conectados, com padrões de disparo, que permitem que a atividade persista mesmo após a remoção dos estímulos externos. A modificação das sinapses entre os neurônios é feita através de regras de aprendizagem *Hebbiana*. Neste trabalho, o modelo *Cell Assemblies* foi adaptado a fim de aprender os relacionamentos entre os termos de uma coleção de documentos. Esses relacionamentos são utilizados para expandir a consulta original com termos relacionados. A avaliação experimental sobre uma coleção de testes padrão em Recuperação de Informações mostrou que algumas consultas melhoraram significativamente seus resultados com a técnica proposta.

**Palavras-chave:** Expansão de consultas, recuperação de informações, redes neurais, aprendizado *hebbiano*.

**Cell Assemblies for Query Expansion**

# ABSTRACT

One of the main tasks in Information Retrieval is to match a user query to the documents that are relevant for it. This matching is challenging because in many cases the keywords the user chooses will be different from the words the authors of the relevant documents have used. Throughout the years, many approaches have been proposed to deal with this problem. One of the most popular consists in expanding the query with related terms with the goal of retrieving more relevant documents. In this work, we propose a new method in which a Cell Assembly model is applied for query expansion. Cell Assemblies are reverberating circuits of neurons that can persist long beyond the initial stimulus has ceased. They learn through Hebbian Learning rules and have been used to simulate the formation and the usage of human concepts. We adapted the Cell Assembly model to learn relationships between the terms in a document collection. These relationships are then used to augment the original queries. Our experiments use standard Information Retrieval test collections and show that some queries significantly improved their results with the proposed technique.

**Keywords:** Query expansion, information retrieval, neural networks, hebbian learning.

# 1 INTRODUÇÃO

Redes Neurais constituem uma área da Ciência da Computação ligada à Inteligência Artificial, com a finalidade de implementar modelos matemáticos que se assemelhem às estruturas neurais biológicas. Dessa forma, apresentam capacidade de adaptar os seus parâmetros como resultado da interação com o meio externo, melhorando gradativamente o seu desempenho na solução de um determinado problema (FERNEDA, 2006). A utilização de técnicas de Redes Neurais no processo de Recuperação de Informações tem como objetivo melhorar a qualidade dos Sistemas de Recuperação de Informações.

Recuperação de Informações trata da representação, armazenamento, organização e acesso aos elementos de informação (BAEZA-YATES; RIBEIRO-NETO, 1999). A consulta feita para satisfazer a necessidade de informação do usuário, normalmente é traduzida em um conjunto de palavras-chave, e submetidas a um Sistemas de Recuperação de Informações (ou um motor de busca), que recupera os itens (documentos de texto, páginas web, imagens, vídeos, etc) que são suscetíveis de satisfazer a necessidade de informação do usuário.

Para a implementação do processo de Recuperação de Informações utilizando um modelo de Redes Neurais, os documentos são descritos por um conjunto de termos representativos desses documentos. E, para cada termo do documento é associado um peso relacionado à correlação entre os termos e o documento. As técnicas mais utilizadas para fazer uma seleção de documentos baseiam-se em métodos estatísticos de Recuperação de Informações.

Um dos principais desafios em encontrar documentos relevantes correspondentes à consulta do usuário é que, em muitos casos, os termos utilizados pelo usuário são diferentes dos termos utilizados pelo autor que escreveu o documento. Dois fenômenos linguísticos contribuem para os maus resultados dos Sistemas de Recuperação de Informações: a sinonímia e a polissemia. A sinonímia refere-se ao fato de o mesmo conceito poder ser expresso por palavras diferentes, como por exemplo *automóvel*, *carro* e *veículo*. A polissemia ocorre quando uma palavra possui vários significados, por exemplo, a palavra *banco* que pode referir-se à instituição financeira, aos "bancos de areia" em um rio, ou ainda, assentos como na expressão "bancos de praça". Ao longo dos anos, várias abordagens têm sido propostas para resolver esses problemas. A Expansão de Consultas é um dos métodos mais utilizados para resolver o problema de sinonímia. A ideia básica é expandir a consulta original com sinônimos e termos relacionados (mais específicos ou mais genéricos), a fim de aumentar o número de documentos relevantes recuperados.

A utilização de Redes Neurais no processo de Recuperação de Informações é uma alternativa natural a ser analisada devido a sua reconhecida capacidade de identificar padrões (BAEZA-YATES; RIBEIRO-NETO, 1999). Assim sendo, esta abordagem também pode ser utilizada para selecionar documentos relevantes.

*Cell Assemblies*, propostas por Hebb (1949), são conjuntos de neurônios interligados que possuem uma grande força sináptica; é uma rede recorrente que pode permanecer ativa mesmo após cessarem os estímulos externos. *Cell Assemblies* representam um conceito ou unidade de pensamento. Um conceito pode ser entendido como um objeto do domínio de interesse. Caso o objeto exista dentro do domínio, ele é considerado um exemplo positivo; caso esse objeto não pertença ao domínio, ele é considerado um exemplo negativo (HAYKIN, 1998). O princípio central diz que as células que estão correlacionadas formam conexões excitatórias sob atividade neural repetitiva. Neurônios que disparam juntos desenvolvem conexões mais fortes.

O modelo CANT (HUYCK, 1999) gera um tipo de *Cell Assemblies* que pode ser utilizada para Recuperação de Informações através da modelagem das relações semânticas entre os termos. Essas relações semânticas são derivadas das características da distribuição dos termos na coleção de documentos. Esse método foi baseado na hipótese de que estatísticas da coocorrência de termos fornecem informações úteis sobre as relações semânticas entre os termos.

O objetivo deste trabalho é propor um novo método de Expansão de Consultas através de um modelo *Cell Assemblies*. Para esse fim, pretende-se adequar o modelo CANT para realizar Recuperação de Informações monolíngue tradicional baseada em texto, para aplicação em grandes volumes de dados. As *Cell Assemblies* devem identificar padrões de relacionamentos entre os termos. Esses relacionamentos são utilizados para expandir a consulta original visando recuperar documentos relevantes para a consulta do usuário.

Os experimentos de validação do modelo utilizaram a coleção *Los Angeles Times*, que contém 113 mil artigos publicados em 1994. Os documentos passaram por etapas de pré-processamento que incluem: *stemming* e remoção de caracteres especiais, valores numéricos e *stopwords*. Os resultados mostram que algumas consultas melhoraram os resultados de maneira significativa com a utilização do método proposto.

Este trabalho está organizado da seguinte maneira:

O Capítulo 2 fornece os principais conceitos de Recuperação de Informações e Redes Neurais Artificiais, áreas de pesquisa em que esta dissertação está inserida.

O Capítulo 3 apresenta os trabalhos relacionados nos tópicos de Expansão de Consultas e Redes Neurais para Recuperação de Informações.

O Capítulo 4 detalha o modelo *Cell Assemblies*, descrevendo suas características e o método proposto.

O Capítulo 5 descreve os experimentos feitos para validar o modelo e a análise dos resultados.

Por fim, o Capítulo 6 apresenta algumas considerações finais, bem como os trabalhos futuros.

# 2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta o contexto no qual esta dissertação está inserida com o intuito de fornecer uma melhor compreensão dos assuntos abordados neste trabalho. São apresentados os principais conceitos e características de cada uma das duas grandes áreas nas quais o método proposto está inserido: Recuperação de Informações (Seção 2.1) e Redes Neurais (Seção 2.2). A Seção 2.3 apresenta uma visão geral do uso de Redes Neurais no processo de Recuperação de Informações. O resumo do capítulo está na Seção 2.4.

## 2.1 Recuperação de Informações

O volume de informações disponibilizado aos usuários é muito maior do que a capacidade desses de encontrar e organizar as que buscam e necessitam. Para diminuir essa dificuldade e, auxiliar o usuário na busca das informações desejadas em grandes repositórios de informações, foram desenvolvidos os Sistemas de Recuperação de Informações (SRI). Segundo Kowalski (1997), um SRI é um sistema de computador capaz de armazenar, recuperar e manter informações. A representação e organização devem possibilitar ao usuário acesso rápido e fácil às informações de seu interesse.

Para Manning et al. (2008), Recuperação de Informações (RI) é uma área da computação que trata do armazenamento de documentos e da recuperação automática de informações a partir deles. Esses autores também definem o papel de um SRI como sendo o de encontrar material (geralmente documentos), de uma natureza não estruturada (geralmente texto), que satisfaça a necessidade de uma informação dentro de grandes coleções (geralmente armazenadas em computadores).

Um típico SRI pré-processa internamente os documentos, a fim de manter o controle dos termos que são utilizados durante o processo de indexação. Essa fase envolve algumas mudanças nos documentos e pode incluir técnicas como (BAEZA-YATES; RIBEIRO-NETO, 1999):

**tokenização** - técnica utilizada para extrair termos do texto. Segundo Manning et al. (2008), *token* é uma sequência de caracteres em um documento. Geralmente, essa etapa descarta caracteres especiais, como pontuação e contrações.

**remoção de *stopwords*** - entendem-se como *stopwords* as palavras extremamente comuns no idioma e consideradas com pouco valor semântico. As listas de *stopwords* geralmente são compostas por conectores linguísticos como: artigos, preposições, conjunções, pronomes, advérbios, etc. Esses termos não participam da indexação.

***stemming* ou radicalização** - processo que reduz um termo a sua raiz morfológica. A raiz de uma palavra é o conjunto de caracteres que está presente em todas as suas derivações. *Stemming* obtém uma representação única para palavras que apontam para um mesmo conceito. Isso permite encontrar mais textos sobre um mesmo assunto sem

a necessidade de usar variações linguísticas, como: plurais, aumentativo, masculino e feminino, etc.

**indexação** - para diminuir o tempo de processamento os SRI realizam uma etapa de catalogação dos documentos, que pode ser atualizada a cada nova inserção ou em intervalos pré-determinados. Essa etapa gera um índice. Uma lista de itens e seus atributos principais compõem a estrutura do índice. E, esses formam a base de todo o SRI (BAEZA-YATES; RIBEIRO-NETO, 1999). Os atributos dos itens descrevem detalhes do item, como o número de documentos em que ele ocorre e sua frequência no documento (HERSH, 2009). Os itens do índice são unidades de informação adequadas para a comparação com os termos da consulta. Podem ser, simplesmente, termos encontrados na coleção de documentos ou termos que representem o conteúdo do documento. Cada item pode ser mais ou menos representativo do texto que está contido em um documento, e essa variação deve ser conhecida. Uma das técnicas de indexação mais utilizadas é o índice invertido. O **índice invertido** possui duas partes principais: ***(i)*** a estrutura da busca, que possui o vocabulário com os termos distintos e, ***(ii)*** os documentos onde a palavra ocorre. As consultas são feitas através dos termos procurados, retornando os documentos onde esses termos aparecem. Essa estrutura está ilustrada na Figura 2.1.

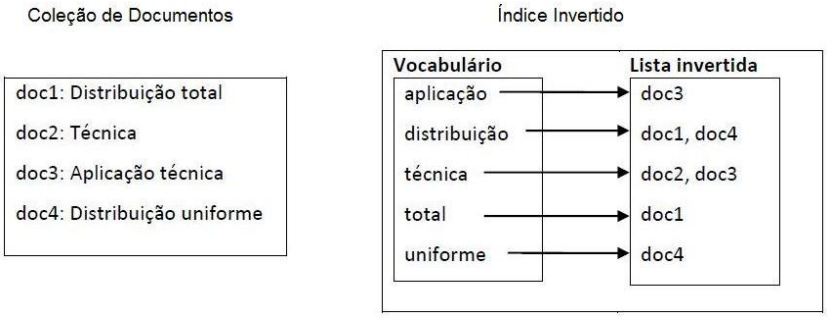



Figura 2.1: Exemplo de formação de índice invertido

A busca indexada consiste em comparar os termos da consulta com os termos do índice e retornar os documentos relevantes à pesquisa ordenados por um critério pré-especificado. Esse critério pode ser, por exemplo, ordem alfabética, cronológica ou por peso dos termos nos documentos. A aplicação de funções de similaridade em sistemas de RI permite a elaboração de uma lista de documentos ordenados de acordo com o seu escore em relação à consulta do usuário (HERSH, 2009).

O esquema ***TF-IDF*** (*Term Frequency - Inverse Document Frequency*) (BAEZA-YATES; RIBEIRO-NETO, 1999) é bastante utilizado em RI e mede, respectivamente, a frequência do termo no documento e a frequência do termo no conjunto de documentos da coleção. Termos que ocorrem com maior frequência em um documento são bons discriminadores do mesmo e, portanto, possuem maior peso. Por outro lado, termos que ocorrem frequentemente no conjunto de documentos não são capazes de diferenciar um documento do outro. Portanto, os melhores termos de indexação são aqueles que aparecem com maior frequência em um documento (alto valor de TF), e aparecem com pouca frequência dentro da coleção (alto valor de IDF).

Um modelo de RI define representações para consultas e documentos e como compará-los. Um típico SRI pré-processa internamente o texto dos documentos (tokenização, remoção de *stopwords* e *stemming*) antes de indexá-lo. Essa etapa é executada de maneira *off-line*. Quando o usuário submete uma consulta, o SRI aplica o mesmo pré-processamento do texto na consulta e, em seguida, o SRI passa a fazer a correspondência

da consulta com os documentos disponíveis. A correspondência é feita através da aplicação de uma função de similaridade, que atribui uma pontuação de similaridade de um documento $d$ em resposta à consulta $q$. Tais resultados são utilizados para gerar a lista ordenada dos documentos que possivelmente atendem à necessidade de informação do usuário. Esse processo está representado na Figura 2.2.

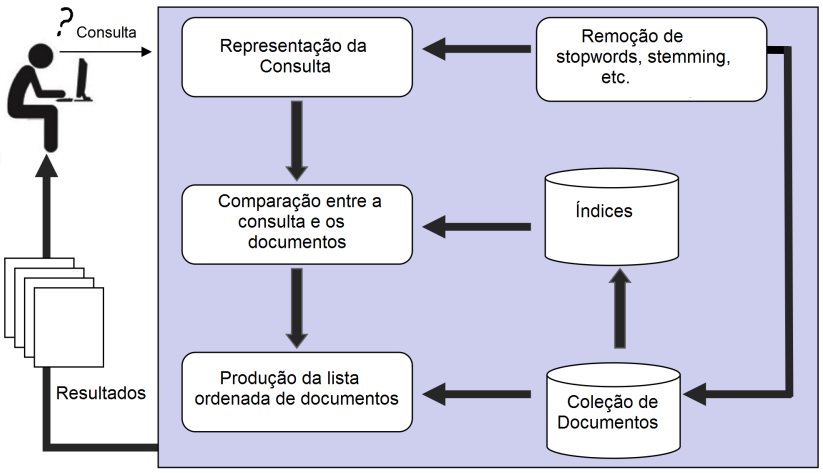



Figura 2.2: Arquitetura típica de um Sistema de Recuperação de Informações

Os SRI utilizam técnicas que visam melhorar os resultados das consultas feitas pelos usuários. Uma das maneiras encontradas é utilizar métodos que expandem automaticamente as consultas dos usuários. A **Expansão de Consultas (EC)** é baseada na correlação dos termos da coleção, e é útil porque a linguagem natural permite que as pessoas utilizem termos e expressões diferentes para indicar um mesmo objeto (XU; CROFT, 1996).

Um SRI requer uma consulta precisa e compreensível para que possa executar a busca e classificação dos documentos e, para que, apenas documentos relevantes sejam apresentados para o usuário. Contudo, as especificações das consultas são limitadas pelo vocabulário do usuário e pelo seu conhecimento do domínio que está sendo consultado. A EC tem como objetivo recuperar os documentos que contêm não apenas os termos da consulta mas, também, os termos que são semanticamente similares a eles. Por exemplo, se a busca do usuário for *"livro"*, documentos que contenham a palavra *"romance"*, não são retornados para o usuário, pois os motores de busca não "entendem" que *"romance"* é um tipo de *"livro"*.

A intenção é possibilitar a recuperação de documentos, mesmo que eles não possuam termos com a mesma grafia dos termos presentes na consulta original. O que diferencia os tipos de EC é o método pelo qual esses termos adicionais são escolhidos. A expansão automática de consultas pode ser classificada como local ou global.

**Métodos locais** utilizam informações do conjunto de documentos recuperados pela consulta original para escolher termos para expansão. Esses métodos, que também são conhecidos como Realimentação de Relevantes (*relevance feedback*), normalmente necessitam da intervenção do usuário para selecionar alguns documentos que ele entende como relevantes. As abordagens que utilizam essa técnica geralmente seguem as seguintes etapas (BILLERBECK; ZOBEL, 2004): ***(i)*** a consulta original é submetida e processada para gerar uma lista inicial de documentos recuperados; ***(ii)*** a partir da lista de documentos recuperados, o usuário indica quais contém informação potencialmente relevante para ele; ***(iii)*** os termos mais relevantes dos documentos escolhidos são adicio-

nados a consulta original, efetuando a expansão das consultas; ***(iv)*** a consulta expandida é processada e os novos resultados são apresentados ao usuário.

Esse processo pode ser repetido por mais de uma iteração, sendo que a cada execução o resultado da busca do usuário deve apresentar melhorias nos resultados. Como vantagem do método, o usuário precisa formular a consulta uma única vez, nas outras iterações ele interage apenas com o sistema abstraindo-se do processo de formulação, simplesmente identificando documentos como relevantes ou não (BILLERBECK; ZOBEL, 2004).

A técnica de *Pseudo-relevance feedback* pode ser vista como uma evolução da técnica de *relevance feedback*, porém não utiliza a participação do usuário. Nela, após o processo inicial de recuperação, assume-se que os $n$ documentos melhor classificados são relevantes, e desses são extraídos termos ou expressões que realimentarão de forma automática a consulta inicial (MANNING; RAGHAVAN; SCHTZE, 2008).

**Métodos globais** não necessitam da intervenção do usuário, a expansão é realizada através do estudo das relações entre os termos em toda a coleção de documentos, dessa forma o processamento da consulta ocorre apenas uma vez, e a análise das relações entre termos da coleção pode ser pré-computada (XU; CROFT, 1996). Esses métodos expandem a consulta sem levar em consideração os resultados recuperados pela consulta original. Isso geralmente é obtido com um *thesaurus* ou WordNet. *Thesaurus* (BAEZA-YATES; RIBEIRO-NETO, 1999) é um dicionário controlado em um dado domínio de conhecimento e é utilizado para identificar sinônimos e entidades linguísticas que são semanticamente semelhantes. Enquanto que, a WordNet (FELLBAUM, 1998) é organizada em conjuntos de sinônimos com termos de mesmo significado, permitindo buscas por nodos semanticamente relacionados.

Para o processo de RI ser executado, é necessária a utilização de um modelo de RI. Os modelos de RI são divididos em dois grandes grupos: ***(i)*** modelos clássicos ou tradicionais de RI e ***(ii)*** modelos alternativos de RI. Os modelos que primeiramente deram sustentação aos SRI foram classificados como sendo os Modelos Clássicos de RI que são: booleano, vetorial e probabilístico. Já os modelos alternativos pretendem agregar ao processo de RI outras técnicas computacionais como Indexação Semântica Latente, Redes Neurais, entre outros (BAEZA-YATES; RIBEIRO-NETO, 1999).

Na literatura, vários modelos alternativos de RI são sugeridos. Esses modelos unem técnicas computacionais na busca por um melhor desempenho na qualidade das informações recuperadas para suprir as necessidades dos usuários. As Redes Neurais, conforme é apresentado na Seção 2.2, podem executar muito bem a tarefa de comparar um dado padrão a um grande número de possíveis padrões e, assim sendo, esta estrutura também pode ser utilizada para selecionar documentos relevantes.

## 2.2 Redes Neurais

Redes Neurais (RN), no contexto da Ciência da Computação, buscam simular o processamento de informações pelo cérebro humano em um modelo computacional. A simulação dessa técnica utiliza sua principal característica que é a capacidade de aprender com novas experiências. Uma RN assemelha-se ao cérebro em dois pontos: ***(i)*** pesos sinápticos são utilizados para armazenar o conhecimento; e ***(ii)*** o conhecimento é obtido através de etapas de aprendizagem.

Sinapse é o nome dado à conexão existente entre os neurônios que formam a RN e, a essas conexões são atribuídos valores, que são chamados de pesos sinápticos (CUNNINGHAM et al., 1997). As RN têm na sua constituição uma série de neurônios artificiais

que são conectados entre si, formando uma rede de elementos de processamento (HAYKIN, 1998). A Figura 2.3 demonstra uma RN como sendo um grafo, onde os neurônios correspondem aos nós do grafo e as arestas que unem os nodos representam as sinapses presentes no cérebro humano.

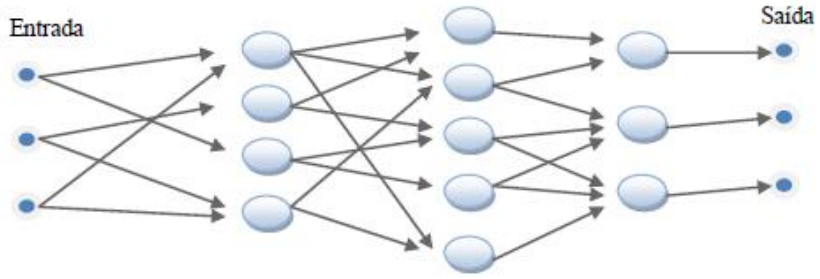



Figura 2.3: Representação de uma RN em forma de um grafo (HAYKIN, 1998)

O tipo de aprendizado é determinado pela técnica empregada no processo de ajuste dos pesos sinápticos (parâmetros da RN). Um conjunto de procedimentos bem definidos recebe o nome de algoritmo de aprendizado. Os algoritmos diferem, basicamente, pela maneira através da qual o ajuste dos pesos é feito. Os diversos métodos para treinamento de RN que têm sido desenvolvidos podem ser agrupados em três paradigmas: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço. Esses tipos de aprendizado distinguem-se pelo conhecimento prévio que se deve possuir sobre o comportamento desejado (HAYKIN, 1998).

**Aprendizado supervisionado**: abastece a rede com exemplos das respostas desejadas; esse conhecimento equivale ao de um professor que é capaz de determinar a saída correta para um conjunto de entradas. O objetivo é ajustar os parâmetros da rede, de forma a encontrar uma ligação entre os pares de entrada e saída fornecidos. Para isto, é necessário ter um conhecimento prévio do comportamento que se deseja, ou se espera da RN. Esta técnica é utilizada para classificação e reconhecimento de padrões, predição (ou previsão) de séries temporais, identificação de sistemas, controle de processos, entre outros.

**Aprendizado não-supervisionado**: não são fornecidos exemplos das respostas desejadas, o aprendizado deriva apenas das observações dos estados do ambiente, sem conhecimento prévio. Não existe um agente externo para acompanhar o processo de aprendizado. A RN processa as entradas e, detectando suas regularidades, tenta progressivamente estabelecer representações internas para codificar características e classificá-las automaticamente. Para que seja possível encontrar padrões é preciso que exista redundância nos dados de entrada. Este paradigma também é utilizado em classificação de padrões. Algumas técnicas bem conhecidas para aprendizado não-supervisionado são: Regra de Hebb e Aprendizado por Competição. A regra de Hebb propõe que conexão sináptica entre dois neurônios seja reforçada sempre que esses estiverem ativos. A ideia geral do aprendizado por competição é que, dado um padrão de entrada, as unidades de saída devem disputar entre si para serem ativadas.

**Aprendizado por reforço**: é uma técnica que possibilita a aprendizagem a partir da interação com o ambiente. A rede opera por tentativa e erro. Durante o processo de

aprendizagem, a rede tenta algumas ações (saídas) e recebe um sinal de reforço (estímulo) do ambiente que permite avaliar a qualidade de sua ação. Aprender por reforço significa aprender o que fazer de modo a maximizar um sinal numérico de recompensa. Não se dispõe da informação sobre quais ações devem ser tomadas, como é o caso no aprendizado supervisionado. O sistema de aprendizagem deve descobrir quais ações tem mais chances de produzir recompensa e realizá-las.

As RN possuem a capacidade de adaptar os seus parâmetros como resultados da interação com o meio externo, melhorando gradualmente o seu desempenho na solução de um determinado problema (HAYKIN, 1998). Segundo Ferneda (2006) a característica mais importante de uma RN é a sua capacidade de aprendizagem por meio de exemplos (treino da RN), essa capacidade faz com que a RN seja capaz de melhorar seu próprio desempenho.

## 2.3 Redes Neurais como modelo para Recuperação de Informações

A utilização de RN no processo de RI é uma alternativa natural a ser analisada devido a sua reconhecida capacidade de generalização, e por possuir um processo de treinamento considerado rápido (BAEZA-YATES; RIBEIRO-NETO, 1999). Vários autores (KWOK, 1989; BAEZA-YATES; RIBEIRO-NETO, 1999; FERNEDA, 2006) associam a tripla {termos da consulta, documentos e termos de índice} que faz parte do processo de RI como sendo uma RN de três camadas (Figura 2.4). O processo de inferência é iniciado pelos termos de consulta que ativam os termos de indexação. Os documentos a serem recuperados recebem sinais dos termos de busca que foram ativados pelos termos de indexação, respectivamente.

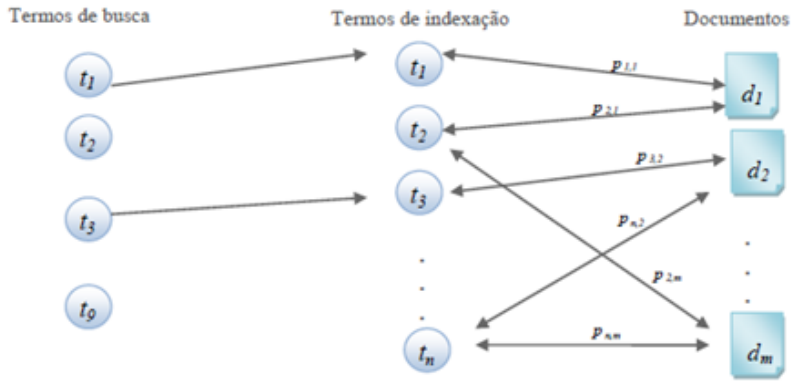



Figura 2.4: Camadas de uma rede neural para RI: uma para os termos da consulta, outra para os termos do documento e a terceira para os documentos (BAEZA-YATES; RIBEIRO-NETO, 1999)

Os estudos para utilização das técnicas de RN em processos de RI começaram em meados dos anos 80. Mozer (1984) utilizou uma arquitetura simples que não dispunha da habilidade de gerar inferências e assim aprender com as experiências, habilidade esta que se caracteriza como sendo uma das principais das RN. Em experimentos com esse modelo (BEIN; SMOLENSKY, 1988), os resultados encontrados foram considerados satisfatórios e percebeu-se que a propagação da ativação através de termos relacionados podem colaborar com o desempenho dos SRI.

Geralmente, pode-se assumir que cada termo é um neurônio, que o documento é

um conjunto de neurônios, e o processo de aprendizagem ocorre através das alterações sinápticas, que podem ser modificadas através de regras de aprendizagem (HUYCK; ORENGO, 2005). O processo, então, segue as seguintes etapas (KWOK, 1989):

- a consulta é submetida pelo usuário;
- os termos de indexação, quando ativados pelos termos de busca, enviam sinais para os documentos; esses sinais são multiplicados pelos pesos de cada ligação;
- os documentos que foram ativados enviam sinais que são conduzidos de volta aos termos de indexação;
- ao receberem esses estímulos, os termos de indexação enviam novos sinais aos documentos e então o processo se repete;
- esse processo se propagará até a estabilização da RN, quando cessam as ativações entre seus nós.

O resultado final da busca será o conjunto dos documentos que foram ativados, cada qual com um nível ativação que pode ser entendido como o grau de relevância do documento em relação à busca do usuário. Ao final do processo de pesquisa, o grau de ativação de cada documento pode ser utilizado como critério de ordenamento dos itens resultantes. Os documentos com maior nível de ativação são, geralmente, aqueles que possuem todos os termos utilizados na expressão de busca, seguidos dos documentos que possuem somente alguns dos termos de busca e dos que foram apenas inferidos durante o processo de pesquisa (HE, 1999).

Dentre as vantagens que este modelo apresenta sobre os modelos clássicos de RI, destacam-se: ***(i)*** o fato de não possuir nenhum programa externo operando sobre a RN, ou seja, é uma rede autoprocessada caracterizada pelo "comportamento inteligente" que é uma característica global em modelos de RN (FERNEDA, 2006; REGGIA; SUTTON G.G., 1988); e ***(ii)*** os modelos de RN exibem comportamentos derivados de interações locais que ocorrem ao mesmo tempo entre os nós da rede através das suas inúmeras conexões sinápticas (HE, 1999).

Além destas, outras vantagens das RN que podem ser úteis ao processo de RI são comentadas a seguir: a capacidade de generalização das RN permite retornar o documento mais próximo, isto é, quando o exato termo da consulta não for encontrado em nenhum documento, é capaz de retornar os documentos com os termos que foram mais ativados pela consulta (HE, 1999). A habilidade das RN de permitir o aprendizado através de exemplos possibilita que toda a vez que um novo padrão for apresentado à rede, o algoritmo de treinamento ajuste os pesos existentes, com objetivo de representar melhor o padrão apresentado. Os processos de apresentação repetitiva e generalização do conjunto de dados permitem que, mesmo após o término do treinamento, a rede tenha condição de reconhecer padrões que nunca foram apresentados a ela (HAYKIN, 1998).

As dificuldades encontradas em experiências com o modelo de RN para RI remetem ao tempo computacional considerável que o processo de treinamento pode demandar, e às ligações que são formadas de maneira aleatória e que não possuem peso após o treino. Estas ligações são consequência do grande número de termos com baixa relevância dentro do documento (MOKRIS; SKOVAJSOVA, 2005).

## 2.4 Sumário

Esse Capítulo apresentou as principais características e conceitos das áreas envolvidas na proposta desta dissertação. Citou as técnicas de pré-processamento dos SRI, o processo de indexação e formação do índice invertido. Métodos de EC como *relevance feedback*, *pseudo-relevance feedback*, *thesaurus*, Wordnet e os tipos de aprendizado das RN são importantes para o entendimentos dos próximos capítulos deste trabalho. Da mesma forma, foi explicada a utilização desses conceitos na integração de RN para RI, e comentadas as vantagens e limitações da utilização do método.

O Capítulo 3 apresenta os trabalhos que utilizam estes conceitos.

# 3 TRABALHOS RELACIONADOS

Este capítulo apresenta os principais trabalhos que estão relacionados aos assuntos tratados na dissertação. O método proposto está inserido em dois grupos distintos: ***(i)*** RN no processo de RI e ***(ii)*** abordagens para EC em RI. O capítulo está organizado da seguinte forma: na Seção 3.1 são apresentadas abordagens de EC para RI e a Seção 3.2 apresenta trabalhos que utilizam modelos de RN em RI. A Seção 3.3 apresenta as considerações finais do capítulo.

## 3.1 Expansão de Consultas para Recuperação de Informações

As técnicas de EC são utilizadas para adicionar termos semanticamente semelhantes aos utilizados na consulta do usuário. Essas técnicas partem do princípio de que a maioria das consultas não representa, adequadamente, as necessidades dos usuários (CARPINETO et al., 2001a; VOORHEES, 1994; XU; CROFT, 1996).

O que diferencia os tipos de EC é o método pelo qual os termos adicionais são escolhidos. A principal questão é como expandir as consultas. As duas abordagens principais são: as que se baseiam no resultado da consulta e as que fazem uso de alguma base de conhecimento. A primeira abordagem é dependente do processo de busca e utiliza *relevance feedback* para a geração de novos termos. Salton & Buckley (1997) relatam experimentos utilizando diferentes técnicas de *relevance feedback*. Os experimentos, que utilizaram cinco coleções diferentes, obtiveram resultados positivos que variaram de 47% (para a coleção CISI) a 160% (para a coleção *Cranfield* [1]). O estudo concluiu que os melhores resultados são obtidos quando:

- as coleções contêm consultas pequenas. O tamanho médio das consultas da *Cranfield* é de 9,2 palavras, enquanto que o tamanho médio das consultas da coleção CISI é de 28,3 palavras;
- quanto pior for o resultado inicial, maior a possibilidade de melhorias;
- são utilizadas coleções técnicas com assuntos específicos, pois são mais adaptáveis ao processo de *relevance feedback*.

Spink (1994) realizou um estudo sobre o comportamento do usuário ao selecionar os termos da consulta, e ao selecionar os termos para a EC. O autor relatou que os termos que aparecem na consulta original são, em geral, os mais escolhidos pelo usuário para expansão dessa consulta. O estudo concluiu que os usuários são capazes de selecionar

[1] coleções disponíveis em: http://www.cs.utk.edu/ lsi/

corretamente os termos para a EC, porém, como o método requer esforço do usuário não é, realmente, utilizado com frequência.

Apesar de obter resultados significativos para EC, os métodos que utilizam *relevance feedback* não são considerados muito robustos, uma vez que são dependentes de que os primeiros documentos recuperados sejam realmente relevantes (XU; CROFT, 2000).

As abordagens que fazem uso de alguma base de conhecimento geralmente utilizam um *thesaurus* ou WordNet. A EC é feita automaticamente através da adição de sinônimos e termos relacionados existentes nos dicionários. A principal vantagem da utilização desses métodos é que eles não necessitam da intervenção do usuário. A principal limitação é o alto custo de construir um *thesaurus* manualmente (MANNING; RAGHAVAN; SCHTZE, 2008; ZHANG; DENG; LI, 2009).

O trabalho de Voorhees (1994) utilizou a Wordnet, como apoio para a expansão dos termos das consultas, e 74200 documentos disponibilizados pela campanha TREC[2]. O conjunto de sinônimos expandidos foi escolhido manualmente (por um ser humano), levando em consideração o contexto da consulta. O autor do artigo assume que, dessa forma, o desempenho obtido pode ser superior a outros mecanismos automáticos que utilizam essa estratégia para EC. Além disso, relata que as consultas mais curtas foram mais beneficiadas com a técnica, e que os melhores resultados advêm de consultas bem formuladas pelo usuário, em relação aos termos expandidos automaticamente.

Grootjen & Weide (2006) e Parapar et al. (2005) utilizam a WordNet como uma fonte de termos adicionais para complementar a consulta do usuário. Os experimentos mostram que a adição de muitos sinônimos pode prejudicar a expansão, além disso, os resultados não apresentaram nenhuma melhoria significativa com a expansão. Mandala et al. (1998) apresenta um estudo detalhado dos motivos de a WordNet não funcionar para EC. Os principais motivos apontados são: a ausência de grande parte dos relacionamentos entre termos; nomes próprios e outros termos que não estão incluídos na WordNet; e a polissemia, onde uma palavra possui mais de um sentido. Recentemente, Bernhard (2010) utilizou *pseudo-relevance feedback* como um método para ajudar na desambiguação dos termos. A técnica é utilizada pois até mesmo as palavras que não possuem mais de um sentido, podem passar a ter após a EC. Os resultados relatam uma pequena melhora em termos de *MAP*.

Na busca de resolver o problema da polissemia, Qiu & Frei (1993) utilizaram um *thesaurus* para realizar a EC. Na abordagem, todas as palavras são tratadas como conceitos. Os termos escolhidos para a expansão são aqueles que coocorrem com maior frequência entre os termos da consulta, com o objetivo de adicionar os termos mais similares ao conceito da consulta. Os resultados foram inferiores aos alcançados com o uso de *revelance feedback*.

A intenção do processo de modificação da consulta é aproximar a consulta dos documentos mais relevantes e a afastar dos não-relevantes. Mais precisamente, ele consiste em modificar a consulta inicial do usuário de acordo com a reação do usuário em relação aos documentos inicialmente recuperados. Essa alteração visa a construção da formulação ideal da consulta que irá selecionar um novo conjunto de documentos relevantes. Desta forma, os SRI devem aprender com as decisões do usuário e, assim, recuperar documentos cada vez mais próximos das suas necessidades.

[2]http://trec.nist.gov/

## 3.2 Redes Neurais aplicadas à Recuperação de Informações

RN fornecem uma representação do conhecimento prático para aplicações de RI, essa Seção apresenta alguns estudos que utilizam RNs como um modelo para RI. Neurônios representam objetos de RI, tais como palavras-chave, referências e trechos dos documentos. As sinapses são implementadas por meio de conexões bi-direcionais ponderadas que representam o nível de relevância das conexões entre os neurônios. Suas propriedades de aprendizagem de *backpropagation* fornecem meios suficientes para o efeito de reformulação da consulta. O algoritmo *bakpropagation* requer a propagação direta do sinal de entrada através da rede, e a retro-propagação do sinal de erro. O erro é calculado comparando-se o resultado obtido ao valor desejado para a resposta (HAYKIN, 1998).

Uma das primeiras tentativas de utilizar RN para RI foi feita por Belew (1989) que propôs o sistema *Adaptative Information Retrieval (AIR)*, o sistema utiliza uma arquitetura de uma RN de três camadas formadas por: termos de indexação, documentos e autores. As conexões são feitas entre os documentos e seus autores e, entre documentos e seus termos de indexação. O sistema empregou *relevance feedback* permitindo ao usuário atribuir graus de relevância aos itens recuperados. O resultado é utilizado na aprendizagem da RN, que modifica os pesos das conexões entre seus nós. Essa interação pode ser vista como um processo contínuo de aprendizagem e adaptação do sistema aos interesses de seus usuários. Os experimentos utilizaram apenas os títulos de 1600 documentos. O resultado foi uma representação do significado consensual de palavras-chave e documentos compartilhados por um grupo de usuários. A desvantagem do sistema é que ele é aplicável apenas em pequenos domínios do conhecimento, onde os usuários possuam interesses comuns.

Kwok (1989) também utilizou uma RN de três camadas. A ideia apresentada é reformular o modelo probabilístico para RI. Consultas, termos de índices e documentos são representados por neurônios em diferentes camadas. O processamento da consulta é feito através da ativação externa dos neurônios presentes na consulta (primeira camada). Esses neurônios propagam a ativação aos termos de índice (na camada média), que por sua vez, propaga a ativação para os documentos da terceira camada. O nível de ativação de cada neurônio do documento é utilizado para fazer o *ranking* dos documentos em relação à consulta. Os experimentos e os resultados não foram relatados no trabalho.

Da mesma forma, Wilkinson & Hingston (1991) utilizaram uma RN de três camadas que representam as consultas, os termos de índice e os termos dos documentos. Esse estudo realizou experimentos padrões de RI em pequenas coleções de teste. O sistema utiliza a medida *TF-IDF* para o cálculo do peso das conexões. O *ranking* dos documentos relevantes recuperados obteve uma melhora de 14% em relação à medição do coseno. No entanto, esse processo de aprendizagem efetuado durante a etapa de consultas apenas permite a modificação do conhecimento no sistema, não existe gestão real da necessidade do usuário.

Boughanem et al. (1998) propuseram, durante a campanha TREC7, o modelo Mercure como uma RN de três camadas (termos da consulta, termos indexados e termos dos documentos da coleção), para executar um *pseudo-relevance feedback*. Teve como objetivo melhorar a modificação automática de consultas baseada na relevância do documento. Os experimentos utilizaram a coleção AP88 e a lista de documentos relevantes, ambas fornecidas pela campanha TREC. O modelo implementa um processo de RI que utiliza o algoritmo de *backpropagation* através das conexões de pesos. As ligações entre as camadas são simétricas e seus pesos são baseados nas medidas *TF-IDF*. A alteração da consulta é baseada na retropropagação da relevância, que consiste em propagar inver-

samente a relevância do documento, partindo da camada de documentos para a camada de termos. O processo possui as seguintes etapas: os termos dos documentos relevantes e não relevantes tem os seus valores (positivos e negativos) processados e propagados para os termos da consulta; os relacionamentos existentes entre os termos da consulta e os documentos são atualizados; a nova consulta é processada gerando uma nova lista de documentos recuperados. O processo foi repetido três vezes. Esse estudo apresentou uma perda de 5% em termos de precisão média quando comparados à execução do *baseline*.

O trabalho de Mokriš & Skovajsova (2005) descreve um modelo de RN para RI em documentos que utilizam textos em linguagem natural. Esse trabalho utiliza um modelo linguístico e uma abordagem conceitual para análise do texto nos documentos. O sistema foi dividido em três módulos: administrador, indexação e usuário. O módulo *Administrador* é responsável pelo processamento do conjunto de documentos. O módulo *Indexação* é encarregado pela criação de índices e palavras-chave. O módulo *Usuário* processa a consulta e busca por documentos relevantes. Esses três subsistemas de RI são a representação de duas RN processadas em três camadas, sendo a primeira: camada de consulta, a segunda: camada de palavras-chave e a terceira: camada de documentos. O processamento da rede se inicia quando o usuário informa a consulta, que é transformada em uma palavra-chave. A RN verifica se esse termo é uma palavra-chave do documento ou não. Caso isso seja verdadeiro, o termo é adicionado ao vetor de palavras-chave. A base do modelo foi desenvolvida no Matlab[3], com 13 palavras-chave e uma coleção com 90 documentos. Depois da realização do treinamento do conjunto de palavras-chave, e feita a associação com os documentos analisados, percebeu-se que somente quando as palavras-chave treinadas foram utilizadas é que houve a recuperação dos documentos relevantes.

Em um estudo preliminar de Huyck & Orengo (2005) foi demonstrado que *Cell Assemblies* podem ser utilizadas para executar categorização e RI. Os experimentos, que utilizaram 1400 documentos da coleção *Cranfield* e 425 documentos da coleção *Time Magazine*, visavam aprender sobre o relacionamento entre as palavras existentes nos documentos propiciando, assim, a recuperação de documentos semelhantes através da rede de relacionamentos formada entre eles. O processamento foi dividido em etapas, sendo que a cada etapa os neurônios com nível de ativação maior que um dado limiar, seriam disparados e os níveis de ativação para todos os neurônios pós-sinápticos atualizados. O cálculo da similaridade entre as consultas e os documentos utilizou a correlação de Pearson[4]. O coeficiente de correlação varia de -1 a 1, sendo que a correlação de 1 significa identificação dos padrões e a correlação -1 sugere que os padrões não compartilham quaisquer características. O *ranking* dos documentos foi decrescentemente ordenado por similaridade. Apesar de considerar o trabalho apenas como exploratório, os autores entenderam os resultados como positivos. Os SRI obtiveram 40% de precisão média na coleção *Time Magazine* e 28% de precisão média na coleção *Cranfield*, sendo que todos os resultados foram semelhantes ao *baseline*.

Desjardins et al. (2006) propuseram uma RN autoassociativa para encontrar correspondências entre consultas e documentos. Nesse tipo de rede, todos os nodos são interconectados a fim de identificar padrões de coocorrência dentro da coleção de documentos, e entre os termos da consulta. A recuperação é baseada na similaridade entre os padrões encontrados nos documentos e os padrões encontrados na consulta. Os autores relatam o experimento utilizando 2000 documentos e 7 consultas, que foram selecionados da coleção FT943, disponibilizada pela campanha TREC. Os pesos sinápticos são ajustados pela

[3]http://www.mathworks.com/products/matlab/

[4]http://pearsoncorrelation.com/

regra *Hebbiana* e os resultados mostram que a RN autoassociativa ultrapassou o VSM em baixos níveis de revocação.

Roberson & Dankel (2007) utilizaram um modelo de Redes Neurais Morfológicas (RNM), que se difere das demais RNs pela maneira que o processo computacional ocorre com os nodos. Multiplicação e adição foram substituídas por adição e máximo (ou mínimo), respectivamente. O trabalho teve como objetivo a criação de um mecanismo de consulta que transformasse as consultas do usuário em uma RNM capaz de filtrar os vetores de documentos em um espaço semântico latente. Os experimentos utilizaram o *corpus* da coleção *Time Magazine*, composto por 425 documentos. Como resultado, a RNM possui um cálculo não-linear antes do disparo. Os experimentos relataram que seus resultados foram significantemente piores que o *Vector Space Model* (VSM). VSM é um modelo algébrico utilizado para representar os documentos e os termos como vetores em um espaço $n$-dimensional, onde $n$ representa a quantidade de termos únicos que ocorrem no interior de todos os documentos. Os autores relataram que os resultados foram abaixo do esperado em termos de precisão, porém comprovaram seu potencial para melhoria de desempenho quando utilizado em um SRI.

Mais recentemente Raiber & Kurland (2010) utililizaram um método de reconhecimento de padrões baseado na classificação através de exemplos, o SVM (*Support Vector Machine*). As consultas são comparadas individualmente a cada um dos documentos. O objetivo do treinamento é agrupar os documentos mais relevantes para cada consulta conforme os resultados obtidos, em termos de *MAP*. Esta classificação é utilizada para realizar um tipo de *relevance feedback* automático, sem a necessidade de intervenção do usuário. Os experimentos utilizaram títulos de consultas, obtidos em coleções disponíveis na campanha TREC, e os cinco documentos relevantes mais bem classificados para cada consulta. Os resultados obtidos foram considerados positivos em termos de *MAP*.

## 3.3 Sumário

Este Capítulo apresentou os trabalhos relacionados ao método proposto, que está detalhado na Seção 4.2. A principal característica da utilização de RN para o processo de RI é a sua capacidade de aprender a cada iteração. A técnica permite atribuir um caráter dinâmico a esses sistemas, dado que os resultados das consultas podem ser reavaliados e alterados, de acordo com a especificação de relevância atribuída pelos usuários aos documentos recuperados.

As abordagens para aplicação da técnica podem variar de acordo com o modelo adotado, contudo, a bibliografia existente refere-se apenas ao potencial para melhoria de desempenho quando aplicado em SRI. A maioria dos experimentos apresentados em livros ou artigos utiliza um ambiente controlado, com um conjunto pequeno de documentos. Tais experimentos priorizam a observação da evolução das ativações, que representam os documentos após um determinado número de iterações. Assim sendo, o desempenho computacional desses modelos em situações reais pode ser considerado ainda desconhecido.

Os trabalhos existentes que utilizam RN para RI, ou não informam sobre os resultados experimentais (BELEW, 1989; KWOK, 1989), ou utilizaram coleções bastante pequenas (RAIBER; KURLAND, 2010; DESJARDINS; PROULX; GODIN, 2006; HUYCK; ORENGO, 2005; ROBERSON; DANKEL, 2007; WILKINSON; HINGSTON, 1991), ou ainda, falharam em produzir melhorias (BOUGHANEM et al., 1998).

Desta forma, vinte anos após os primeiros estudos serem publicados, a aplicação de

RN para RI ainda permanece uma questão em aberto. Neste trabalho pretende-se ir além, no sentido de encontrar uma solução através da aplicação de um tipo diferente de RN para RI, com a realização de experimentos em uma coleção de testes padrão, utilizada em avaliações na área de RI.

# 4 *CELL ASSEMBLIES* PARA RECUPERAÇÃO DE INFORMAÇÕES

RN são conhecidos modelos computacionais inspirados no funcionamento neural humano. Contudo, sistemas naturais funcionam de maneira diferente da maioria das RN. O modelo CANT (*Connections, Associations and Network Technology)* foi utilizado neste trabalho por gerar um tipo de *Cell Assemblies* que pode ser utilizado para EC (HUYCK, 1999). A principal contribuição deste trabalho é a proposta de um novo método de EC baseada no modelo CA. Este capítulo descreve o modelo *Cell Assemblies* (Seção 4.1) e a aplicação do modelo para EC (Seção 4.2). Um resumo do capítulo está descrito na Seção 4.3.

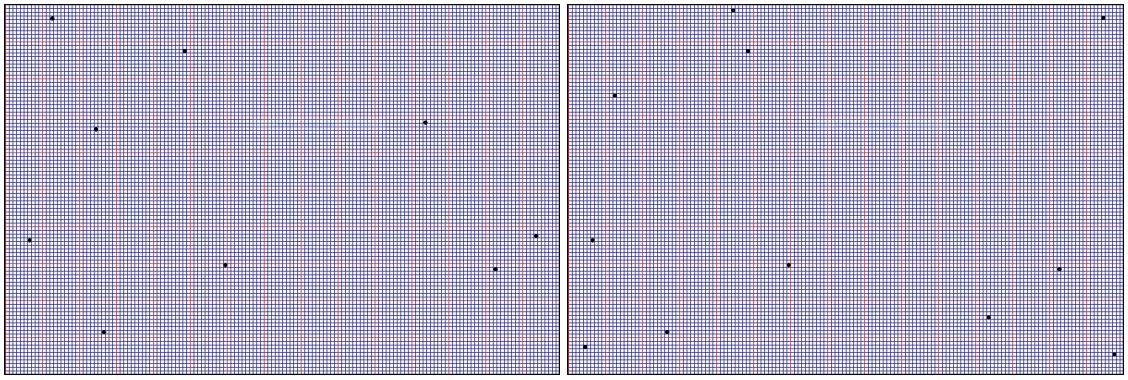
## 4.1 Modelo *Cell Assemblies*

*Cell Assemblies* (CA) ou conjunto neuronal, proposto por Hebb (1949), sugere que grupos de neurônios ativos no cérebro, são responsáveis pelo armazenamento do conhecimento dos seres humanos. CA são grupos de neurônios que possuem uma força sináptica substancial. Através da propagação da ativação, as CA podem permanecer ativas mesmo depois de terminados os estímulos externos (HUYCK, 1999). Recentemente o modelo CA (HUYCK, 2004; IVANCICH; HUYCK; KAPLAN, 1999) tem sido utilizado para resolver outros problemas como RI (HUYCK; ORENGO, 2005), categorização (HUYCK, 2007), agentes inteligentes (HUYCK; BYRNE, 2009), entre outras aplicações. Esta Seção descreve a arquitetura do modelo CA que foi utilizada no método proposto.

O neurônio é a base do modelo CA. O neurônio possui axônios (conexões), um valor de ativação e um limiar de ativação que deve ser alcançado antes de disparar. Os neurônios são conectados a outros neurônios por um pequeno número de sinapses (ligações).

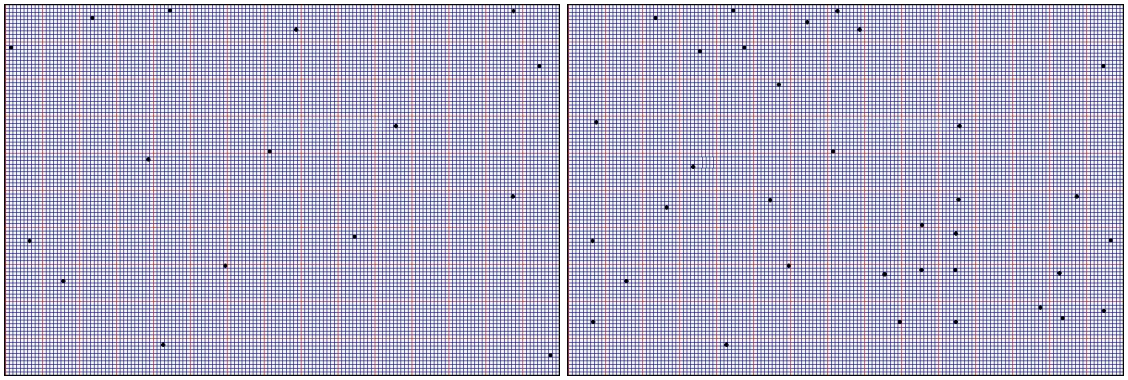
O processo é dividido em discretos passos de tempo. Em cada passo de tempo, dispara o neurônio que estiver com um nível de ativação maior do que um dado limiar. E o nível de ativação, para todos os neurônios pré e pós-sinápticos, é atualizado. Os neurônios que compõe a RN são randomicamente conectados, via sinapses, a outros 40 neurônios (HUYCK; ORENGO, 2005).

As imagens da Figura 4.1 ilustram a rede como uma matriz retangular, onde cada célula representa cada um dos neurônios da coleção, ordenados alfabeticamente. As células em destaque representam os neurônios disparando. No passo de tempo 1 (*a*) os neurônios ativos correspondem aos termos da consulta original. Nos passos de tempo 2 (*b*), 3 (*c*) e 4 (*d*), a ativação da rede é espalhada para os neurônios correlacionados até o passo de tempo 5 (*e*), quando o estado da rede é salvo e os termos que correspondem aos neurônios que estavam disparando são identificados e utilizados para expandir a consulta original.



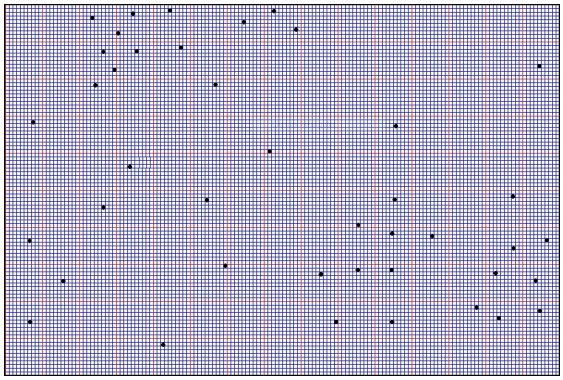
(a) Passo de tempo 1

(b) Passo de tempo 2



(c) Passo de tempo 3

(d) Passo de tempo 4



(e) Passo de tempo 5

Figura 4.1: Exemplo da ativação da rede

O modelo foi criado para operar de maneira similar a de um sistema neural natural. A ideia é baseada em Hebb (1949) que estabelece que uma CA é equivalente a um conceito. A cada execução a rede deve ativar uma determinada CA. O modelo CANT utiliza um tipo de fadiga dos neurônios chamada *fatiging Leaky Integrate and Fire* (fLIF) que está brevemente descrita na Subseção 4.1.1 (HUYCK, 2007).

### 4.1.1 Neurônios

A rede CA é baseada em neurônios fLIF. Os neurônios recebem ativação de outros neurônios via conexões sinápticas. Um neurônio dispara apenas quando possuir ativação suficiente que ultrapasse um determinado limiar de ativação. Quando o neurônio dispara, ele envia ativação proporcional ao peso sináptico, através das suas conexões a outros neurônios. Ao disparar, o neurônio perde toda a sua energia, porém quando o neurônio não dispara, ele perde uma fração de sua energia. Os neurônios são chamados de pré e pós-sinápticos e a energia flui do neurônio pré-sináptico para o neurônio pós-sináptico.

O modelo CANT gera CA a partir de um conjunto de parâmetros. Os neurônios possuem cinco propriedades principais, sendo que as três primeiras são bastante comuns em modelos de RN, e as duas últimas propriedades são mais raras:

1. Força da Conexão (*Connection Strength*) - os neurônios possuem conexões com outros neurônios. As conexões podem ter força positiva ou negativa e a ativação contínua é simulada através de passos de tempo.

2. Ativação (*Activation*) - um neurônio possui um nível de ativação que é baseado, em grande parte, no comportamento do disparo dos neurônios que estão conectados a ele.

3. Limiar de Ativação (*Activation Threshold*) - um neurônio dispara se possuir ativação suficiente para ultrapassar um determinado limiar de ativação.

4. Decréscimo (*Decay*) - se um neurônio não dispara, ele retém sua atividade mas tem seu valor diminuído. Uma nova entrada pode levar a um ganho de ativação. A constante decréscimo é aplicada a neurônios ativos e inativos.

5. Fadiga (*Fatigue*) - quanto mais tempo um neurônio ficar ativo, maior será o seu limiar de ativação. Os neurônios fadigam e assim é menos provável que disparem.

A ativação (propriedade 2) que os neurônios possuem a cada passo de tempo utiliza o peso das sinapses (propriedade 1), e é descrita pela Equação 4.1 (PASSMORE; HUYCK, 2008). Se conseguir energia suficiente, o neurônio dispara e propaga sua ativação (propriedade 3). Caso não consiga energia suficiente, perde uma porção de sua ativação através da constante de decréscimo (propriedade 4).

$$h_i^t = \frac{h_i^{t-1}}{d} + \sum_{j \in V} w_{ji}, d > 1 \tag{4.1}$$

A ativação atual $h$ de um neurônio $i$ em um dado passo de tempo $t$, é a ativação do último passo de tempo dividida pelo fator de decréscimo $d$, somada à nova ativação. Essa nova ativação é a soma das entradas de todos os neurônios $j \in V$, ponderada pelo valor da sinapse do neurônio $j$ ao neurônio $i$; $V$ sendo o conjunto de todos os neurônios que estão conectados a $i$ e que dispararam em um passo de tempo $(t-1)$. Caso um neurônio tenha disparado em um passo de tempo anterior $(t-1)$ ele não repassa a sua ativação para $t$.

Um neurônio não deve ser capaz de disparar continuamente. Por esse motivo, o modelo utiliza uma fadiga neural (propriedade 5). Essa propriedade é modelada através do aumento do limiar de ativação cada vez que um neurônio é disparado. Quando um neurônio não dispara, seu limiar de ativação diminui a cada passo de tempo até que ele atinja um dado nível de base.

### 4.1.2 Aprendizado *Hebbiano*

O modelo CA utiliza a regra de aprendizado *Hebbiano*. A regra geral do aprendizado *Hebbiano* diz que: caso dois neurônios disparem ao mesmo passo de tempo, a força de sua sinapse deve ser aumentada. Para que não exista saturação da sinapse, o modelo também utiliza uma regra anti-*Hebbiana* que diminui o peso da sinapse a cada vez que o neurônio pré-sináptico disparar e o pós-sináptico não disparar.

Para tratar RI, uma regra de aprendizagem compensatória foi utilizada. Essa regra limita a força sináptica total de um neurônio. Como resultado, os neurônios com poucas correlações têm a sua influência aumentada e neurônios com muitas correlações têm a sua influência reduzida. Essa regra é similar ao IDF, comumente utilizado em RI (HUYCK, 2007).

## 4.2 *Cell Assemblies* aplicadas à Expansão de Consultas

A principal motivação para utilizar CA para EC é explorar a propagação da ativação para expandir a consulta original com termos relacionados. O modelo computacional é baseado no processamento neural dos mamíferos. Os neurônios são conectados unidirecionalmente via sinapses, e o aprendizado ocorre através de modificações sinápticas. As regras de aprendizado não-supervisionado alteram as sinapses, e essas mudanças são baseadas apenas nas propriedades dos neurônios pré e pós-sinápticos.

A rede CA fornece a EC, pois, ao estimular os termos da consulta esses enviam a ativação para os termos relacionados, automaticamente expandindo a consulta através da base de conhecimento formada. O processo de construção automática da base de conhecimento é realizado através de cálculos estatísticos da coocorrência de pares de palavras.

A hipótese que está por trás dessa estratégia é que se duas palavras aparecem próximas em vários documentos, então elas possuem certo relacionamento. O resultado desse processo é um conjunto de conceitos, representados por grupos de palavras que caracterizam uma ideia contida nos documentos da coleção. Esses conceitos são integrados à rede semântica que compõe a base de conhecimento. Essa rede semântica é utilizada para melhorar a eficiência do sistema e recuperar documentos relevantes à consulta do usuário.

A fim de adaptar o modelo de CA para realizar EC, foi necessário desconsiderar algumas limitações que são importantes na modelagem do cérebro de mamíferos. Por exemplo, no cérebro os neurônios provavelmente permanecem conectados aos neurônios que estão mais próximos e não aos mais distantes. Essa restrição é ignorada nas simulações. Porém, a diferença mais importante é que cada termo é representado por um único neurônio, o que não é biologicamente plausível uma vez que os termos são representados por muitos neurônios em várias áreas do cérebro.

No modelo CA, cada termo é representado por um neurônio, e os documentos são representados pelo conjunto de neurônios associados aos termos contidos no documento. Como cada termo é representado por um neurônio, o tamanho da rede é muito reduzido permitindo a codificação de vários milhares de termos, com uma rede de milhares de neurônios. Os documentos são indexados e confrontados com a tarefa de comparar a consulta com o conjunto dos termos.

Para cada termo que ocorre em mais de um documento é atribuído um neurônio. Cada neurônio é conectado a outros neurônios, que representam os termos com os quais ele coocorre pelo menos uma vez em toda a coleção de documentos. A seleção dos termos presentes nas conexões é feita aleatoriamente.

### 4.2.1 Fase de Treinamento

O treinamento da rede é a primeira fase do processo, nesta fase os documentos da coleção são apresentados para a rede CA. Os neurônios correspondentes aos termos presentes nos documentos recebem ativação externa. Seguindo o padrão utilizado no modelo, o formato de representação da rede é um documento. Esse documento contém: ***(i)*** as dimensões da matriz (linhas *x* colunas) que formam a rede, localizadas nas duas primeiras linhas; ***(ii)*** os neurônios, cada qual identificado por seu índice; ***(iii)*** os axônios que estão conectados a cada neurônio, identificados por seus índices e pelo nome da rede; e ***(iv)*** o peso das conexões entre o neurônio e cada axônio conectado a ele. Esse formato está demonstrado na Figura 4.2.

```
307          ⇒ Tamanho da rede
307
0 Neuron
40 Axons
BaseNet  2123  0.1        ⇒ Peso inicial
BaseNet  2235  0.1
.
.
1 Neuron        ⇒ Termo
40 Axons
BaseNet  2498  0.1
BaseNet  9511  0.1
BaseNet  7476  0.1
.
.
.
```

Figura 4.2: Ilustração do formato da rede CA

Conforme os documentos são apresentados ao sistema, os pesos das sinapses entre os neurônios pré e pós-sinápticos são ajustados. O resultado desse ajuste são os novos pesos que devem refletir as relações de coocorrência entre os termos dos documentos. Um novo arquivo é criado, com os pesos já atualizados, conforme ilustrado na Figura 4.3. O processo de aprendizagem ocorre por modificação sináptica (HUYCK; ORENGO, 2005).

```
307          ⇒ Tamanho da rede
307
0 Neuron
40 Axons
BaseNet  2123  0.854        ⇒ Peso ajustado
BaseNet  2235  0.623
.
.
1 Neuron        ⇒ Termo
40 Axons
BaseNet  2498  0.990
BaseNet  9511  0.748
BaseNet  7476  0.672
.
.
.
```

Figura 4.3: Rede CA com os pesos ajustados

Ao final do processo de aprendizagem, são modeladas as relações semânticas entre os termos. Essas relações são formadas de acordo com a distribuição dos termos na coleção de documentos.

### 4.2.2 Fase das Consultas

As consultas consistem de um conjunto de palavras-chave que descrevem a necessidade de informação do usuário. Durante essa fase, as consultas são apresentadas para a rede CA. Os neurônios que representam os termos da consulta recebem estímulos externos. Como resultado, eles disparam enviando ativação para outros neurônios através de conexões sinápticas.

Os estímulos externos e a propagação de ativação continuam por alguns ciclos, e então, o estado da rede é salvo. Esse processo tem o efeito de expandir a consulta inicial com os termos correlacionados. O raciocínio é que, adicionando termos correlatos, o SRI vai recuperar documentos mais relevantes. Para um melhor entendimento, um exemplo de uma consulta expandida está na Tabela 4.1 onde, a consulta original está representada pelos seus radicais, e na terceira coluna estão os termos relacionados, adicionados pela EC.

Tabela 4.1: Exemplo de EC

| **Consulta** | **Radicais dos termos** | **EC** |
|---|---|---|
| C140 | *cellular, develop, docum, industry, mobil, phone, prospect, relev* | *bellsouth, motorola, technolog, telecommun, telephon* |

## 4.3 Sumário

Este Capítulo apresentou o Modelo CA utilizado para RI. A rede CA é baseada em neurônios simples que estão conectados via sinapses. O modelo utiliza uma regra compensatória, semelhante ao IDF, em conjunto com o aprendizado *Hebbiano*. Os objetivos do método são: estudar as características de coocorrência entre os termos da coleção e expandir a consulta do usuário através da inclusão de termos relacionados.

No Capítulo 5 estão descritos os experimentos realizados para validação do método proposto.

# 5 EXPERIMENTOS

Este capítulo apresenta os experimentos realizados que visam avaliar a utilização do método proposto. O modelo CA é empregado como um meio para expandir as consultas dos usuários. O intuito não é modelar o comportamento neural de um especialista humano em recuperação textual, e sim usar o modelo neural de CA como um mecanismo para RI. Com o intuito de avaliar a técnica, foram executados experimentos em três diferentes alternativas de treinamento:

- A coleção LA Times completa;
- Alguns documentos relevantes da coleção LA Times;
- Tópicos individuais de consulta da LA Times.

O ambiente de trabalho está descrito na Seção 5.1, os treinamentos estão explicados na Seção 5.2. Na Seção 5.3, está a análise dos resultados obtidos com os experimentos.

## 5.1 Materiais e Métodos

A Subseção 5.1.1 descreve a Plataforma de Trabalho e a Coleção de Teste utilizada está apresentada na Subseção 5.1.2. Para melhor clareza, também estão descritos: os Tópicos de Consulta na Subseção 5.1.3, e as Métricas de Avaliação na Subseção 5.1.4.

### 5.1.1 Plataforma de Trabalho

Os experimentos foram executados em um microcomputador com processador Intel Core2 Duo, 2GB de memória RAM, 2,40 GHz e 150 GB de disco rígido. Os algoritmos de pré-processamento para criação da matriz de coocorrência e criação da rede de neurônios foram implementados utilizando a linguagem de programação C. Os algoritmos restantes foram implementados na linguagem Java 1.6. Todos os experimentos rodaram em sistema operacional *Windows Vista Business* 32 Bits. Para remover os sufixos das palavras foi utilizado Porter *stemmer* (PORTER, 1980). As *stopwords* foram removidas de acordo com a lista fornecida por SMART[1].

O SRI utilizado foi o Zettair[2], por ser um sistema compacto e rápido. O sistema é considerado muito eficiente para utilização com grande quantidade de dados. O Zettair foi desenvolvido pela RMIT *University* (Austrália) possuindo diversas métricas de RI para definir a similaridade entre consultas e documentos.

[1] ftp://ftp.cs.cornell.edu/pub/smart/english.stop
[2] http://www.seg.rmit.edu.au/zettair/

Neste trabalho utilizou-se a métrica Okapi BM25 (JONES; WALKER; ROBERTSON, 2000) que apresentou os melhores resultados em experimentos preliminares. A partir de uma consulta $Q$ contendo um conjunto de palavras-chave $q1, ...q_n$, escore BM25 de um documento $d$ é dada pela Equação 5.1.

$$BM25(d,Q) = \sum_{i=1}^{n} IDF(q_i) \frac{tf(q_i,d) \cdot (k_1+1)}{tf(q_i,d) + k_1(1 - b + b\frac{|D|}{avgdl})} \tag{5.1}$$

De acordo com a equação, $tf(q_i, d)$ é a frequência do termo $q_i$ no documento $d$, $|D|$ é o tamanho (em palavras) do documento $d$; $avgdl$ é a média do tamanho dos documentos na coleção; $k_1$ e $b$ são parâmetros que convencionam a importância de cada termo da consulta e a quantidade de documentos, respectivamente; $IDF(q_i) = \log \frac{N}{n(q_i)}$, onde $N$ é o número total de documentos na coleção e $n(q_i)$ é o número de documentos que contenham $q_i$. Para os experimentos definiu-se $k_1 = 1,2$ e $b = 0,75$.

Os resultados foram avaliados pelo Trec_Eval[3], que é uma ferramenta padrão utilizada pela campanha TREC para avaliar execuções *ad hoc* na área de RI. O arquivo contendo os resultados foi comparado aos conjunto dos documentos julgados relevantes. Para cada consulta, foram recuperados os 1000 documentos com escore (BM25) mais alto.

### 5.1.2 Coleção de Teste

Para validação dos experimentos foi utilizada a coleção do Jornal *Los Angeles Times*, disponibilizada pela campanha CLEF [4], assim como os 50 tópicos das consultas utilizadas. A coleção *Los Angeles Times* (LA Times) é composta por 113005 artigos de jornal, de variados temas, publicados no ano de 1994. A coleção possui 94027 termos distintos e 821 documentos relevantes. A coleção é escrita em inglês e está disponível no formato SGML. A Tabela 5.1 exibe o exemplo de um documento da coleção LA Times. Cada documento contém um número de documento, um id, um título e um texto.

Tabela 5.1: Exemplo de um documento.

| |
|---|
| <DOC> |
| <DOCNO> LA012394-0072 </DOCNO> <DOCID> 006347 </DOCID> |
| <HEADLINE>TIME TO CARE FOR ROSES </HEADLINE> |
| <TEXT> Give attention to roses at this time. Prune |
| them before spraying to reduce the total area that needs |
| to be sprayed at this time. Remember to drench the |
| branches and trunk until the material runs off the branches. |
| </TEXT> |
| < /DOC> |

### 5.1.3 Tópicos de Consulta

Foram utilizados os tópicos de consulta, numerados de 91 a 140, do ano de 2002. Os tópicos de consulta são compostos por um número de identificação, um título, uma descrição e uma narrativa. A Tabela 5.2 exemplifica um tópico de consulta. Como é feito normalmente em RI, as perguntas utilizadas foram compostas por termos do título, da descrição e da narração.

[3] http://trec.nist.gov/trec_eval/

[4] http://www.clef-campaign.org/

Tabela 5.2: Exemplo de um tópico de consulta.

| |
|---|
| <top> |
| <num> C140 < /num> |
| <EN-title> Mobile phones < /EN-title> |
| <EN-desc> Prospects for the use of cellular phones. < /EN-desc> |
| <EN-narr> Relevant documents report on the prospects for the use of cellular phones and the development of the mobile phone industry. < /EN-narr> |
| < /top> |

Todos os documentos e as consultas necessitaram de um pré-processamento que consiste na remoção de pontuações, normalização dos caracteres em minúsculo, remoção de *stopwords* e sufixos dos termos (*stemming*), além da formatação dos documentos para indexação no sistema Zettair.

### 5.1.4 Métricas de Avaliação

Os experimentos realizados utilizam as seguintes métricas para avaliação dos resultados:

- Precisão (*precision*) (MANNING; RAGHAVAN; SCHTZE, 2008) - mensura a fração de documentos relevantes que são corretamente recuperados, para cada consulta.

- Revocação (*recall*) (MANNING; RAGHAVAN; SCHTZE, 2008) - mensura a fração de todos os documentos relevantes da coleção que são retornados com a utilização do modelo.

  Revocação e precisão são definidas de acordo com as equações:

  $$\textit{Precisão (P)} = \frac{Total\ de\ Relevantes\ Recuperados}{Total\ de\ Recuperados} \tag{5.2}$$

  e,

  $$\textit{Revocação (R)} = \frac{Total\ de\ Relevantes\ Recuperados}{Total\ de\ Relevantes} \tag{5.3}$$

  Onde, *Relevantes* é o conjunto de documentos relevantes, e *Recuperados* é o conjunto dos documentos retornados pelo modelo CA, utilizados para a expansão das consultas.

- Teste T (WILLIAM, 2006) - avalia se as médias de dois grupos são estatisticamente diferentes.

- Curva de Precisão $x$ Revocação (MANNING; RAGHAVAN; SCHTZE, 2008) - gráfico que exibe a evolução da precisão em função da revocação. Avalia o comportamento do sistema, através da posição do documento retornado. Destacam-se os sistemas que retornam documentos relevantes no topo do *ranking*, pois esses devem ter maior importância. Nessa representação, a precisão média para cada consulta é calculada tomando como referência 11 pontos no intervalo [0, 1], representando os níveis de revocação.

Por exemplo, a precisão média no nível 0,2 constitui a precisão aferida nos resultados após a análise de 20% dos documentos relevantes retornados (a partir do topo do *ranking*). Dessa forma, se o sistema apresenta alta precisão nos primeiros níveis de revocação, significa que é alta a densidade de documentos relevantes no topo do *ranking*.

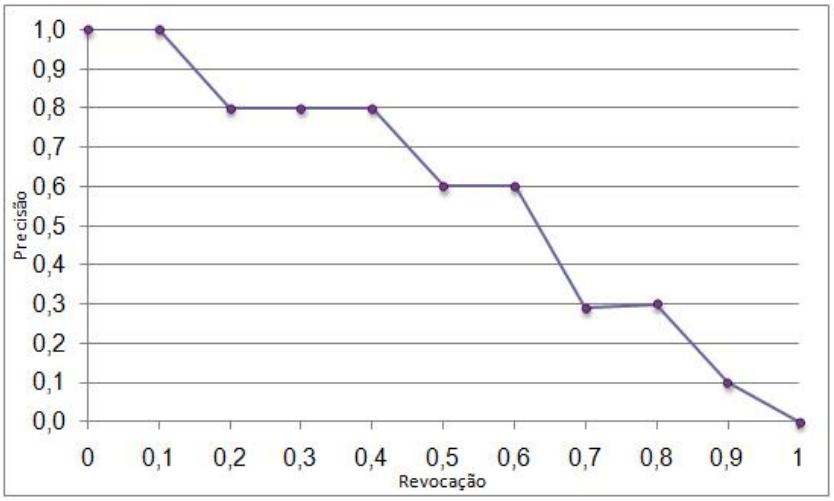



Figura 5.1: Exemplo de curva de precisão média em 11 pontos de revocação

O gráfico da Figura 5.1 representa a curva de precisão-revocação para uma consulta, porém, geralmente os SRI executam diversas consultas. Assim, para gerar um gráfico que sumarize as informações de várias consultas é calculada a média das precisões em cada ponto dessa curva.

- Média das precisões médias (*mean average precision - MAP*) (MANNING; RAGHAVAN; SCHTZE, 2008) - esta é a medida mais utilizada para avaliar resultados de consultas em RI. A MAP enfatiza os documentos retornados no topo do *ranking* e é calculada de acordo com a Equação 5.4.

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk}) \tag{5.4}$$

Onde $|Q|$ corresponde ao número de consultas; $R_{jk}$ é o conjunto de documentos recuperados, ordenados do primeiro até o documento $d_k$; $m_j$ é o número de documentos relevantes para cada consulta $j$; e $P$ representa a precisão dos documentos relevantes recuperados.

## 5.2 Alternativas para Treinamento da Rede Neural

Esta Seção descreve os experimentos realizados e os resultados obtidos. Foram executados 3 tipos de experimentos, que são diferenciados pelo tipo dos neurônios que compõe a rede, e pela quantidade de neurônios envolvidos na fase de treinamento.

O processo de EC é composto das seguintes etapas: durante a apresentação das consultas, os neurônios correspondentes aos termos originais são externamente estimulados. Esta ativação permanece durante cinco ciclos e então o estado da rede é salvo. Os termos

que correspondem aos neurônios que estavam ativados após esses cinco ciclos de ativação são adicionados a consulta original.

A construção da rede CA foi feita através da conexão de cada neurônio a outros 40 neurônios randomicamente selecionados. Os neurônios deveriam coocorrer ao menos em um documento. Todos os neurônios partiram com um peso inicial definido em 0,1. O formato da rede é um documento contendo o tamanho da rede, a identificação do neurônio, a quantidade de axônios, a identificação da rede, a identificação dos axônios e seus respectivos pesos.

Para o treinamento da rede CA, cada documento foi apresentado à rede durante um ciclo. Esse procedimento foi repetido 20 vezes. Ao final desse processo os pesos sinápticos foram ajustados.

Com a finalidade de comparar os resultados obtidos, os experimentos foram executados em duas implementações: uma utilizando CA e EC e outra em que os documentos não sofreram alterações. Ambas implementações utilizaram o sistema Zettair para indexação e avaliação. Para clareza do texto, as execuções que não incluem EC são tratadas como *baseline*.

A Subseção 5.2.1 descreve o experimento onde a RN foi treinada utilizando todos os documentos da coleção. Na Subseção 5.2.2 está o experimento em que o treinamento utilizou alguns documentos relevantes da coleção. A Subseção 5.2.3 descreve o experimento onde o treinamento foi realizado com os tópicos de consulta individuais.

### 5.2.1 Treinamento com a coleção completa

Os experimentos dessa fase foram executados conforme as etapas que estão descritas abaixo:

*(i)* Para o *baseline*, todos os documentos foram indexados no Zettair e os tópicos originais apresentados como consultas. Os processos de EC, criação da rede e treinamento seguiram conforme descritos na Seção 5.2. A definição dos parâmetros da RN foi feita através de heurísticas e de múltiplas tentativas em busca dos melhores resultados. Para a fase de treinamento da rede CA com a coleção completa, os parâmetros foram definidos conforme segue:

- Força da Conexão: 0,06
- Limiar de Ativação: 0,8
- Decréscimo: 2,0
- Fadiga: 0,2

*(ii)* Na fase de avaliação, os neurônios correspondentes aos termos da consulta original foram externamente estimulados. Esta ativação permaneceu por cinco ciclos e, então o estado da rede foi salvo. Os termos correspondentes aos neurônios que estavam ativos foram adicionados a consulta original.

*(iii)* Os resultados para esta execução estão resumidos na Tabela 5.3, que apresenta os resultados obtidos pelo *baseline* e pela EC via CA. O menor número de documentos relevantes recuperados com a EC deve-se a grande quantidade de termos que foram adicionados a consulta original.

As curvas de Precisão-Revocação, exibidas na Figura 5.2, mostram que a EC foi melhor em cinco pontos de revocação. Esse resultado significa que os documentos relevantes

Tabela 5.3: Resultados da execução com todos os documentos

| | ***Baseline*** | **EC via CA** |
|---|---|---|
| MAP | 37,41% | 37,84% |
| Relevantes Recuperados | 505 | 492 |

conseguiram uma melhor pontuação dentro dos documentos recuperados. Para este treinamento foi calculado um Teste-T, que mostrou que não há diferença significativa em relação à baseline (p-value = 0,8494).

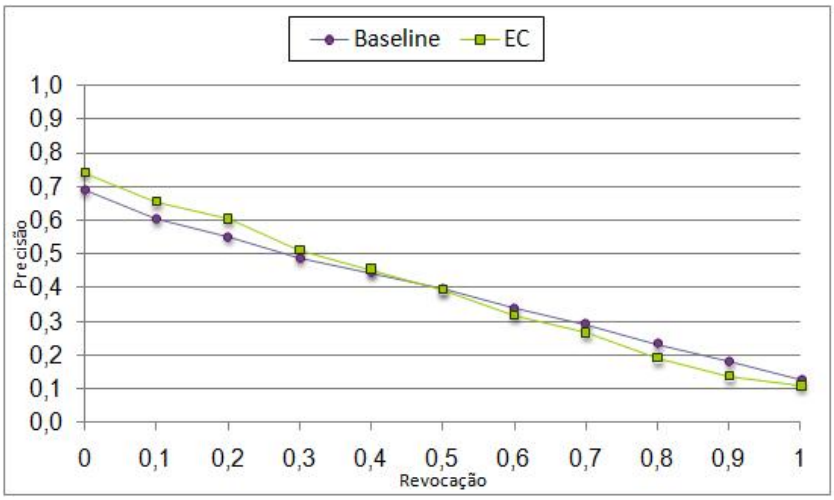



Figura 5.2: Curva de Precisão-Revocação LA Times

Em função das melhorias serem apenas marginais, foi executada uma análise tópico-por-tópico onde foi possível verificar quais tópicos melhoraram e quais pioraram com a EC e CA. O resultado da análise revelou que 12 tópicos melhoraram, 30 tópicos pioraram, e em 8 tópicos não houve diferença no resultado. A Tabela 5.4 mostra os cinco tópicos que mais melhoraram e os que mais pioraram com a utilização de CA para EC, em termos de MAP.

Tabela 5.4: Análise tópico-por-tópico da LA Times

| **Tópicos com Ganho** | | | | **Tópicos com Perda** | | | |
|---|---|---|---|---|---|---|---|
| **Tópico** | ***Baseline*** | **EC via CA** | **Ganho** | **Tópico** | ***Baseline*** | **EC via CA** | **Perda** |
| 100 | 0.1540 | 0.6438 | 49% | 136 | 1.000 | 0.5000 | 50% |
| 91 | 0.2564 | 0.6429 | 39 % | 116 | 0.3911 | 0.2010 | 19 % |
| 139 | 0.0957 | 0.3846 | 29 % | 103 | 0.5298 | 0.3551 | 17 % |
| 129 | 0.3214 | 0.4035 | 8 % | 105 | 0.4373 | 0.2971 | 14 % |
| 133 | 0.0907 | 0.1672 | 8 % | 124 | 0.3940 | 0.3171 | 8 % |

Atribui-se esse ganho em qualidade à adição de poucos termos similares aos termos das consultas originais. Nessa análise preliminar do uso de CA para EC, quando utilizada a coleção LA Times completa, obteve-se um resultado muito próximo ao *baseline*. Com o intuito de examinar o impacto de outros tipos de treinamento utilizando CA, foram feitas execuções apenas com documentos relevantes da coleção, conforme descrito na Subseção 5.2.2.

### 5.2.2 Treinamento com documentos relevantes

Nesta Subseção está descrito o experimento com alguns documentos relevantes da coleção. Para a fase de treinamento foram aleatoriamente escolhidos 398 entre os 821 documentos relevantes. Essa seleção teve como objetivo executar o treinamento apenas com exemplos positivos.

*(i) Baseline*: os documentos que participaram do treinamento não participaram da fase de avaliação e não foram indexados no Zettair. Agiu-se dessa forma com a intenção de não beneficiar esses documentos.

*(ii)* O processo de EC assim como a fase de treinamento seguiram conforme descrito na Seção 5.2. Para a construção da rede CA foram utilizados apenas os neurônios que representam os termos existentes nos documentos relevantes que participaram do treinamento. Nesta execução, os parâmetros utilizados para a fase de treinamento foram os seguintes:

- Força da Conexão: 1,2
- Limiar de Ativação: 0,3
- Decréscimo: 2,0
- Fadiga: 0,6

*(iii)* Resultados: Para a avaliação da série de execuções para EC em comparação ao *baseline*, foram utilizadas as medidas de qualidade descritas na Seção 5.1.4. A Tabela 5.5 mostra um resumo dos resultados em termos de MAP. As curvas de Precisão e Revocação da Figura 5.3 mostram que as execuções para EC com CAs foram superiores ao *baseline* em um nível baixo de revocação ($\leq 0.3$). Isto significa que a EC permitiu a recuperação de documentos relevantes no início do *ranking*. O Teste-T verificou que não há diferença significativa em relação à baseline (p-value = 0,3670).

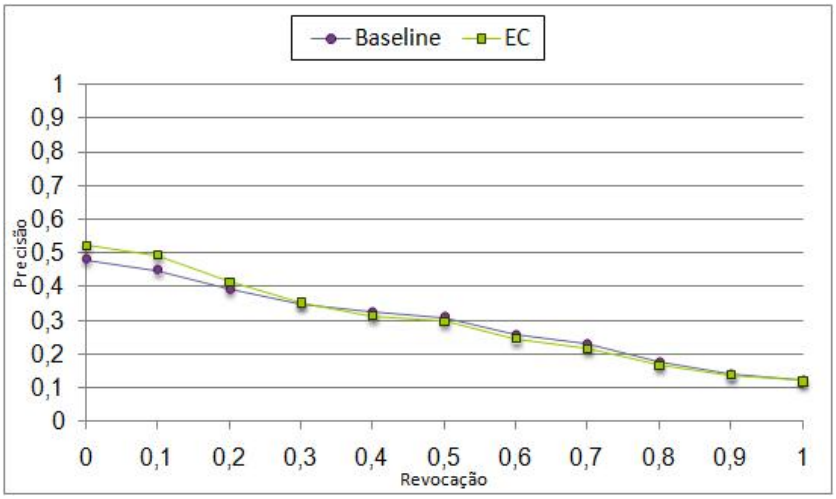



Figura 5.3: Curvas de Precisão-Revocação para Relevantes

Em virtude do ganho alcançado com a EC não ter sido expressivo (2,36%), em termos de MAP, foi realizada uma análise particular em cada tópico. Essa análise tópico-por-tópico também pretendeu verificar como alguns tópicos de consulta melhoraram e outros pioraram com a EC via CAs. O resultado dessa análise mostrou que 20 tópicos de consulta melhoraram, 17 pioraram e 13 permaneceram inalterados. Também mostrou que

Tabela 5.5: Treinamento com Relevantes

| | **Baseline** | **EC via CAs** |
|---|---|---|
| MAP | 27.80% | 28.46% |
| Relevantes Recuperados | 372 | 360 |
| Média de termos na consulta | 18.84 | 29.52 |

as consultas que obtiveram maiores índices de melhora foram as que possuíam poucos documentos relevantes.

Na Tabela 5.6 está a proporção de aumento ou perda da EC em relação ao *baseline*, dos 10 tópicos que obtiveram um maior ganho e dos 10 tópicos com maior perda. Observou-se que os percentuais de melhoria foram maiores que os de perda.

Tabela 5.6: Análise dos tópicos em termos de *MAP*

| **Tópicos que mais se beneficiaram** | | | | |
|---|---|---|---|---|
| **Tópico** | ***Baseline*** | **EC via CA** | **Alteração** | **% Alteração** |
| 129 | 0.1063 | 0.2463 | +0.14 | +132% |
| 126 | 0.0593 | 0.1272 | +0.07 | +115% |
| 133 | 0.1563 | 0.2837 | +0.13 | +82% |
| 94 | 0.2829 | 0.4554 | +0.17 | +61% |
| 91 | 0.2418 | 0.3038 | +0.06 | +26% |
| 121 | 0.4809 | 0.5950 | +0.11 | +24% |
| 140 | 0.1874 | 0.2262 | +0.04 | +21% |
| 99 | 0.0663 | 0.0799 | +0.01 | +21% |
| 135 | 0.0970 | 0.1100 | +0.01 | +13% |
| 131 | 0.2906 | 0.3215 | +0.03 | +11% |
| **Tópicos com maiores perdas** | | | | |
| **Tópicos** | ***Baseline*** | **EC via CA** | **Alteração** | **% Alteração** |
| 116 | 0.3097 | 0.1615 | -0.15 | -92% |
| 122 | 0.1224 | 0.0770 | -0.05 | -59% |
| 124 | 0.2477 | 0.1691 | -0.08 | -46% |
| 95 | 0.3317 | 0.2775 | -0.05 | -20% |
| 103 | 0.4403 | 0.3855 | -0.05 | -14% |
| 114 | 0.4072 | 0.3784 | -0.03 | -8% |
| 123 | 0.3459 | 0.3243 | -0.02 | -7% |
| 119 | 0.7712 | 0.7374 | -0.03 | -5% |
| 120 | 0.3916 | 0.3771 | -0.01 | -4% |
| 92 | 0.7013 | 0.6876 | -0.01 | -2% |

Nessa execução, o tópico 129 foi o que obteve um maior percentual de ganho. Para esse tópico, os documentos relevantes recuperados na execução para o baseline estavam nas posições de ranking: 3, 22, 103, 109 e 121. Quando aplicada a EC, os documentos relevantes ficaram nas posições: 1, 20, 71, 87 e 116. Percebeu-se que, através da EC os documentos relevantes foram recuperados antes, o que resultou na melhoria do resultado para a consulta.

Em todas as consultas em que o *baseline* obteve um resultado perfeito na recuperação (MAP = 1), a EC também obteve o mesmo resultado. O tópico com maior percentual de

perda foi o 116. A EC falhou em recuperar um dos documentos relevantes. Além disso, os relevantes que foram recuperados estavam em posições inferiores, quando comparados à execução do *baseline*.

Uma das hipóteses de a EC beneficiar algumas consultas enquanto prejudica outras, está no fato de que pretendeu-se ensinar à rede CA muitos conceitos, de uma só vez. Tendo em mente que a coleção de documentos utilizada possui um ano de reportagens de jornal, percebe-se que nela existe uma gama muito ampla de assuntos, incluindo política, esportes, ciências, cultura e entretenimento. Caso esta hipótese fosse verdadeira, ensinando poucos conceitos de cada vez, obtería-se melhores resultados. A fim de testar esta hipótese, foram executados experimentos individuais com os tópicos. Esse experimento está descrito na Subseção 5.2.3.

### 5.2.3 Treinamento com tópicos individuais

Com o intuito de ensinar poucos conceitos de cada vez à rede CA, foram escolhidos 10 tópicos de consultas. Desses, cinco estavam entre os que mais se beneficiaram com a EC via CA e cinco estavam entre os que pioraram.

Nesta etapa, utilizou-se o processo de construção da rede CA, treinamento e avaliação, conforme já descrito na Seção 5.2. Para o treinamento, utilizou-se os mesmos parâmetros descritos na Seção 5.2.2. Os resultados estão ilustrados na Tabela 5.7. A Figura 5.4 mostra os gráficos, com as curvas de Precisão e Revocação, demonstrando que quase todos os tópicos que melhoraram com o treinamento em grupo, melhoraram ainda mais quando treinados individualmente.

Tabela 5.7: Precisão média para tópicos treinados individualmente

| **Tópicos com Ganho** | | | | | |
|---|---|---|---|---|---|
| **Tópicos** | ***Baseline*** | **EC** | **EC-Ind.** | **Alteração** | **% Alteração** |
| 129 | 0.1063 | 0.2463 | 0.6624 | 0.56 | 523% |
| 133 | 0.1563 | 0.2837 | 0.4299 | 0.27 | 175% |
| 94 | 0.2829 | 0.4554 | 0.6002 | 0.32 | 112% |
| 91 | 0.2418 | 0.3038 | 0.2963 | 0.05 | 23% |
| 140 | 0.1874 | 0.2262 | 0.2266 | 0.04 | 21% |
| **MAP** | 0.1949 | 0.3032 | 0.4431 | 0.25 | 127% |
| **Tópicos com Perdas** | | | | | |
| **Tópicos** | ***Baseline*** | **EC** | **EC-Ind.** | **Alteração** | **% Alteração** |
| 114 | 0.4072 | 0.3784 | 0.4435 | 0.04 | 9% |
| 116 | 0.3097 | 0.1615 | 0.3294 | 0.02 | 6% |
| 103 | 0.4403 | 0.3855 | 0.3020 | -0.14 | -31% |
| 119 | 0.7712 | 0.7374 | 0.6006 | -0.17 | -22% |
| 95 | 0.3317 | 0.2775 | 0.3184 | -0.01 | -4% |
| **MAP** | 0.4520 | 0.3881 | 0.3988 | -0.05 | -12% |

O tópico 129 obteve uma melhora substancial, pois obteve um resultado de 10% no *baseline* e alcançou 66%, em relação ao MAP, com treinamento individual. O treinamento individual também beneficiou os tópicos 114 e 116 que haviam piorado com o treinamento em grupo. Em RI, uma diferença proporcional de mais de 5% é considerada notável e uma diferença de mais de 10% é considerada importante (BUCKLEY; VOORHEES, 2000).

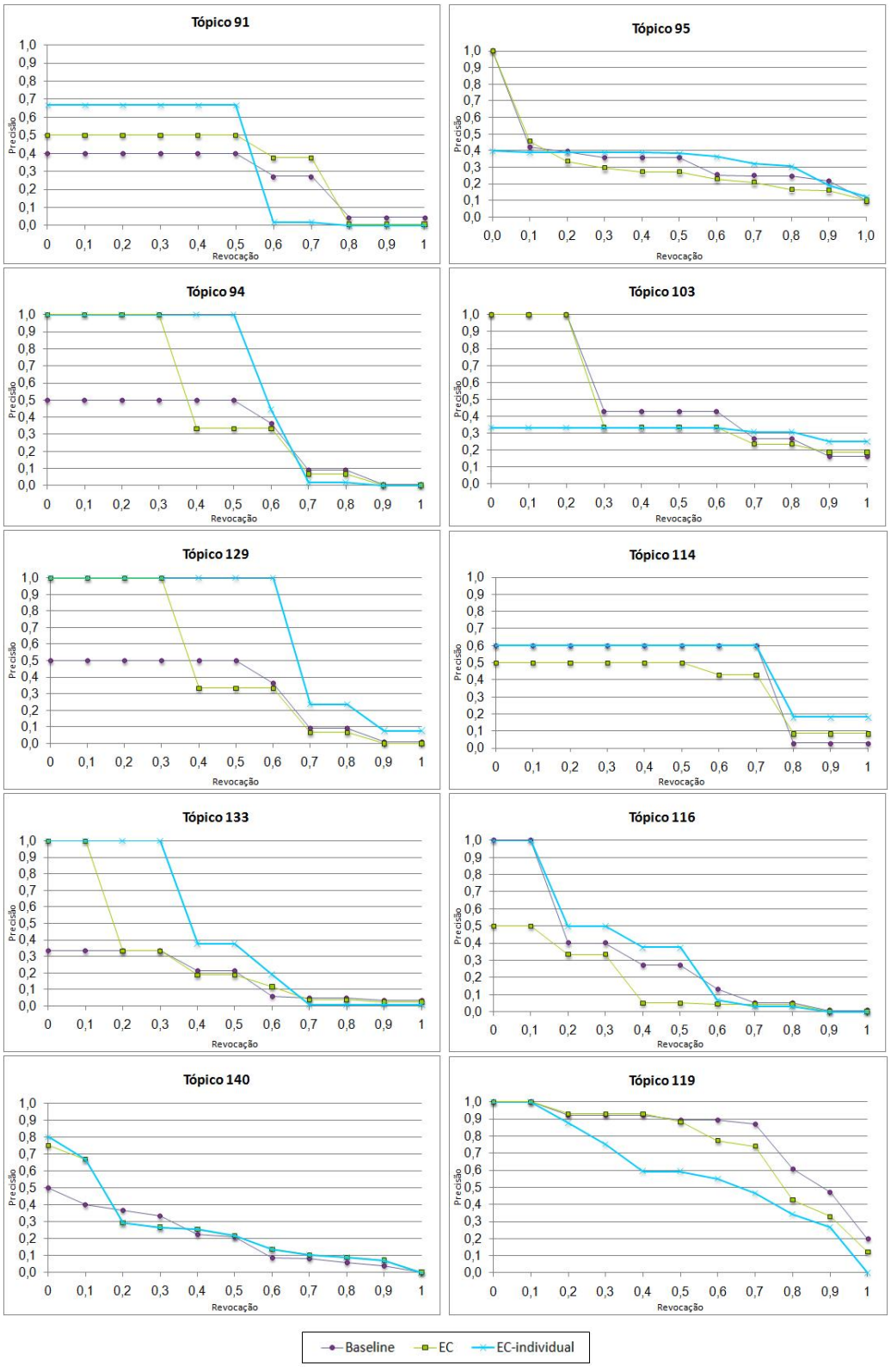



Figura 5.4: Curvas de Precisão-Revocação para os tópicos individuais. A primeira coluna contém os tópicos que melhoraram com EC via CA, a segunda coluna contém os tópicos que pioraram

Estas melhorias sustentam a hipótese de que quando ensinamos poucos conceitos obtemos melhores resultados. Entretanto, as consultas 95, 103 e 119 permaneceram com

os resultados menores que os obtidos no *baseline*, mesmo com treinamento individual. Uma análise mais profunda mostrou que esses tópicos possuem termos muito frequentes na coleção. Na consulta 103, por exemplo, o termo mais raro está presente em 448 documentos. Os outros dois tópicos seguem esta mesma tendência. Nesses casos, a expansão não poderia beneficiar a consulta visto que os termos adicionados acabam por recuperar ainda mais documentos, o que pode diminuir ainda mais o escore dos documentos relevantes. Em contraste, consultas com termos mais raros obtiveram melhores resultados. Por exemplo, o tópico 94 mais do que duplicou o seu resultado em termos de MAP utilizando a EC. Esse tópico possui termos que aparecem em apenas 31 documentos dentro da coleção completa.

## 5.3 Discussão

As conclusões encontradas, através dos experimentos feitos, ratificam descobertas feitas em outros estudos. Em RI, diferentes consultas respondem melhor a diferentes abordagens. Mandl & Womser-Hacker (2002) demonstraram esse fato quando avaliaram várias execuções da campanha CLEF. Eles observaram um desvio padrão elevado para a performance dos tópicos e um desvio padrão elevado para a performance de cada execução. Nesse estudo, sua conclusão foi a de que nenhuma execução tem um bom desempenho em todos os tópicos. Na mesma linha, Orengo & Huyck (2006) mostraram que a principal fonte de impacto sobre a mudança no desempenho produzido por Realimentação de Relevantes (*Relevance Feedback*) são os tópicos. No entanto, os autores não relataram as características dos tópicos que os fez reagir de maneira diferente. No presente trabalho identificou-se que, a frequência dos termos da consulta nos documentos é uma característica desse tipo.

Além disso, vários estudos analisam o impacto dos termos utilizados para a EC (DRAGONI; DA COSTA PEREIRA; TETTAMANZI, 2010; CAO et al., 2008; CUI et al., 2002; CARPINETO et al., 2001b). Esses trabalhos mostraram que a técnica de EC pode ter efeito negativo sobre o sistema, diminuindo a precisão (a fração dos documentos recuperados que são relevantes) caso não haja uma escolha adequada dos termos utilizados para EC.

Não houve meios de se fazer uma comparação direta, entre os resultados obtidos com a utilização CA para EC, e outros modelos que aplicam RN para RI, tendo em vista que as coleções de testes utilizadas são diferentes. Assim mesmo, vale a pena relatar seus resultados. Boughanem et al. (1998) reportaram uma perda em termos de MAP quando utiliza a coleção TREC. Sua *baseline* obteve 27% enquanto sua abordagem obteve 22,78%. Roberson & Dankel (2007) também reportaram perdas em comparação ao VSM com a coleção *Time*. Desjardins et al. (2006) obtiveram maior precisão com baixos níveis de revocação (menor que 0,3). Contudo, o estudo utilizou apenas 7 consultas e a coleção continha apenas 2 mil documentos. Similarmente, Huyck & Orengo (2005) reportam melhoria com a coleção *Cranfield*, que também é muito pequena.

Durante o processo de treinamento, alguns fatores podem influenciar nos resultados obtidos com o método. O tempo necessário para o treinamento com a coleção completa representou uma limitação importante. Da mesma maneira, a definição dos parâmetros se revelou uma tarefa pouco trivial, sendo que várias execuções foram necessárias até a melhor combinação de parâmetros ser encontrada.

Apesar das limitações encontradas, e considerando os resultados encontrados nos outros estudos já comentados, o método proposto neste trabalho obteve um bom resultado.

Na média, o percentual de melhora foi modesto, mas algumas consultas mais do que dobraram seus escores.

# 6 CONCLUSÃO

Esta dissertação apresenta um método de EC que utiliza modelo CA. O fundamento é modelar os relacionamentos entre os termos e então utilizar esses relacionamentos para expandir a consulta original através dos termos relacionados. O método foi avaliado utilizando a metodologia padrão de RI, para os experimentos utilizou uma coleção com 113 mil documentos e 50 tópicos de consultas.

O trabalho detalha o modelo CA utilizado para o desenvolvimento do método assim como descreve os experimentos. Foram testados 3 tipos de treinamento, variando o tipo dos neurônios que formam a rede, e a quantidade de neurônios envolvidos na fase de treinamento. A principal contribuição deste trabalho é identificar padrões de relacionamento entre os termos, a fim de expandir a consulta do usuário com termos semanticamente relacionados.

Os resultados obtidos mostraram uma melhora pequena (2,36%), em termos de MAP. Contudo, algumas consultas obtiveram melhorias significativas e mais do que dobraram os resultados. Em uma análise mais profunda dos resultados percebeu-se que as características dos tópicos de consultas afetam o resultado da expansão. Tópicos com termos muito frequentes na coleção de documentos, tendem a piorar com a expansão, enquanto tópicos com termos mais raros tendem a melhorar.

Como produção científica foi publicado o artigo:

(VOLPE; MOREIRA, 2010) **Cell Assemblies for Query Expansion**. In: WORKSHOP DE TESES E DISSERTAÇÕES EM BANCOS DE DADOS, WTDBD, 9., SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, SBBD, 25. Anais... Belo Horizonte: SBC, 2010. Este artigo apresenta uma visão geral do modelo CA para Expansão de Consultas.

VOLPE, I.; MOREIRA, V.; HUYCK C. **Cell Assemblies for Query Expansion in Information Retrieval**, aceito no INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS 2011- IJCNN'11. Este artigo descreve o método proposto e apresenta experimentos realizados.

Neste trabalho, a EC baseou-se apenas na coocorrência entre os termos, sem levar em consideração vários parâmetros que influenciam nos resultados da técnica tais como: a complexidade da consulta, o número de documentos selecionados e o número de termos que devem ser utilizados na expansão. Esses aspectos podem ser passíveis de melhorias em trabalhos futuros, melhorando o modelo atual com o acréscimo de novas técnicas.

A complexidade das consultas pode ser analisada através dos *logs* das consultas. Os buscadores acumulam uma grande quantidade de *logs* contendo a consulta e os documentos que o usuário selecionou. Através dessa análise é possível compreender a consulta e

o tipo de informação que o usuário busca.

A qualidade dos termos utilizados para expansão pode ser analisada através da elaboração de um *ranking* dos termos sugeridos pelo sistema. Os primeiros da lista seriam os termos com maior coocorrência com os termos da consulta. Por fim, a quantidade de termos pode ser verificada através de heurísticas, como no estudo de Cui (2002) onde os primeiros 60 termos foram utilizados para EC.

Além disso, pretende-se fazer um estudo das características dos documentos nas diferentes seções da coleção de notícias. O estudo de Cool et al. (1993) indica que existem vários fatores associados à utilidade ou relevância de documentos, sendo que alguns destes fatores podem ser utilizados melhorar a recuperação da informação desejada.

# REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. [S.l.]: ACM Press / Addison-Wesley, 1999.

BEIN, J.; SMOLENSKY, P. Application of the interactive activation model to document retrieval. In: NEURONIMES 88, Nimes, France. **Proceedings...** [S.l.: s.n.], 1988. p.295–308.

BELEW, R. K. Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. **SIGIR Forum - Special Interest Group on Information Retrieval Forum**, [S.l.], v.23, p.11–20, 1989.

BERNHARD, D. Query expansion based on pseudo relevance feedback from definition clusters. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS: POSTERS, 23., Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2010. p.54–62. (COLING '10).

BILLERBECK, B.; ZOBEL, J. Techniques for efficient query expansion. In: STRING PROCESSING AND INFORMATION RETRIEVAL SYMP. **Proceedings...** Springer-Verlag, 2004. p.30–42.

BOUGHANEM, M. et al. Mercure At Trec7. In: TREC-7. **Anais...** [S.l.: s.n.], 1998.

BUCKLEY, C.; VOORHEES, E. M. Evaluating evaluation measure stability. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23., New York, NY, USA. **Proceedings...** ACM, 2000. p.33–40. (SIGIR '00).

CAO, G. et al. Selecting good expansion terms for pseudo-relevance feedback. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 31., New York, NY, USA. **Proceedings...** ACM, 2008. p.243–250. (SIGIR '08).

CARPINETO, C. et al. An information-theoretic approach to automatic query expansion. **ACM Trans. Inf. Syst.**, New York, NY, USA, v.19, p.1–27, January 2001.

CARPINETO, C. et al. An information-theoretic approach to automatic query expansion. **ACM Trans. Inf. Syst.**, New York, NY, USA, v.19, p.1–27, January 2001.

COOL, C. et al. Characteristics of texts affecting relevance judgments. In: NATIONAL ONLINE MEETING, 14. **Proceedings...** [S.l.: s.n.], 1993. p.77–84.

CUI, H. et al. Probabilistic query expansion using query logs. In: WORLD WIDE WEB, 11., New York, NY, USA. **Proceedings...** ACM, 2002. p.325–332. (WWW '02).

CUNNINGHAM, S. J. et al. Applying Connectionist Models to Information Retrieval. **Cybernetics & Human Knowing**, [S.l.], v.8, p.64–74, 1997.

DESJARDINS, G.; PROULX, R.; GODIN, R. An Auto-Associative Neural Network for Information Retrieval. In: NEURAL NETWORKS, 2006. IJCNN '06. INTERNATIONAL JOINT CONFERENCE ON. **Anais...** [S.l.: s.n.], 2006. p.3492 –3498.

DRAGONI, M.; DA COSTA PEREIRA, C.; TETTAMANZI, A. G. B. An ontological representation of documents and queries for information retrieval systems. In: INDUSTRIAL ENGINEERING AND OTHER APPLICATIONS OF APPLIED INTELLIGENT SYSTEMS - VOLUME PART II, 23., Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2010. p.555–564. (IEA/AIE'10).

FELLBAUM, C. **WordNet - An Eletronic Lexical Database. MIT Press**. 1998.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, [S.l.], v.35, p.25–30, 2006.

GROOTJEN, F. A.; WEIDE, T. P. V. d. Conceptual query expansion. **Data Knowl. Eng.**, [S.l.], v.56, n.2, p.174–193, 2006.

HAYKIN, S. **Neural Networks**: a comprehensive foundation. 2nd.ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

HE, Q. **Neural Network and its Application in Information Retrieval**. 1999.

HEBB, D. **The Organization of Behaviour**: a neuropsycological theory. New York: Ed. Wiley, 1949.

HERSH, W. **Information Retrieval**: a health and biomedical perspective. 3.ed. [S.l.]: Springer, 2009. (Health Informatics).

HUYCK, C. **Modelling Cell Assemblies**. 1999. n.ISSN 1462-0871 CS-07.

HUYCK, C. Overlapping cell assemblies from correlators. **Neurocomputing**, [S.l.], v.Volume 56, p.435–439, 2004.

HUYCK, C.; BYRNE, E. **CABot1**: technical report. 2009.

HUYCK, C.; ORENGO, V. Information Retrieval and Categorisation using a Cell Assembly Network. **Neural computing & applications**, [S.l.], v.14, n.4, p.282–289, 2005.

HUYCK, C. R. Creating hierarchical categories using cell assemblies. **Connect. Sci**, [S.l.], v.19, n.1, p.1–24, 2007. 1392501.

IVANCICH, J. E.; HUYCK, C. R.; KAPLAN, S. Cell assemblies as building blocks of larger cognitive structures. **Behavioral and Brain Sciences**, [S.l.], p.pp. 292–293, 1999. 10.1017/S0140525X99331824.

JONES, K. S.; WALKER, S.; ROBERTSON, S. E. A probabilistic model of information retrieval: development and comparative experiments. **Inf. Process. Manage.**, [S.l.], v.36, p.779–808, November 2000.

KOWALSKI, G. **Information Retrieval Systems**: theory and implementation. 1st.ed. Norwell, MA, USA: Kluwer Academic Publishers, 1997.

KWOK, K. L. A neural network for probabilistic information retrieval. **SIGIR Forum - Special Interest Group on Information Retrieval Forum**, [S.l.], v.23, n.SI, p.21–30, 1989. 75338.

MANDALA, R.; TAKENOBU, T.; HOZUMI, T. **The Use of WordNet in Information Retrieval**. 1998.

MANDL, T.; WOMSER-HACKER, C. Linguistic and Statistical Analysis of the CLEF Topics. In: CLEF. **Anais...** Springer, 2002. p.505–511. (Lecture Notes in Computer Science, v.2785).

MANNING, C. D.; RAGHAVAN, P.; SCHTZE, H. **Introduction to Information Retrieval**. [S.l.]: Cambridge University Press, 2008. 1394399.

MOKRIS, I.; SKOVAJSOVA, L. Development of Neural Network Information Retrieval System from Text Documents. **3rd Slovakian Hungarian Joint Symposium on Applied Machine Intelligence**, [S.l.], p.123–131, 2005.

MOZER, M. C. **Inductive Information Retrieval Using Parallel Distributed Computation**. [S.l.]: La Jolla: University of California, San Diego, Institute for Cognitive Science, 1984. (ICS Technical Report 8406).

ORENGO, V. M.; HUYCK, C. R. Relevance feedback and cross-language information retrieval. **Information Processing & Management**, [S.l.], v.42, n.5, p.1203–1217, 2006.

PARAPAR, D.; BARREIRO, A.; LOSADA, D. E. Query expansion using wordnet with a logical model of information retrieval. In: IADIS. **Proceedings...** [S.l.: s.n.], 2005. p.487–494.

PASSMORE, P.; HUYCK, C. Models of cell assembly decay. In: CYBERNETIC INTELLIGENT SYSTEMS, 2008. CIS 2008. 7TH IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2008. p.1 –6.

PORTER, M. An Algorithm for Suffix Stripping. **Program**, [S.l.], 1980.

QIU, Y.; FREI, H.-P. Concept based query expansion. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 16., New York, NY, USA. **Proceedings...** ACM, 1993. p.160–169. (SIGIR '93).

RAIBER, F.; KURLAND, O. On identifying representative relevant documents. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 19., New York, NY, USA. **Proceedings...** ACM, 2010. p.99–108. (CIKM '10).

REGGIA, J.; SUTTON G.G., I. Self-processing networks and their biomedical implications. **Proceedings of the IEEE**, [S.l.], v.76, n.6, p.680 –692, June 1988.

ROBERSON, C.; DANKEL, D. D. I. A Morphological Neural Network Approach to Information Retrieval. In: FLAIRS CONFERENCE. **Anais...** [S.l.: s.n.], 2007. p.184–185.

SALTON, G.; BUCKLEY, C. **Improving retrieval performance by relevance feedback**. [S.l.]: Morgan Kaufmann Publishers Inc., 1997. 355-364p.

SPINK, A. Term Relevance Feedback and Query Expansion: relation to design. In: SIGIR - SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL FORUM. **Anais...** [S.l.: s.n.], 1994. p.81–90.

VOLPE, I.; MOREIRA, V. Cell Assemblies for Query Expansion. In: WORKSHOP DE TESES E DISSERTAÇÕES EM BANCOS DE DADOS, WTDBD, 9., SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, SBBD, 25. **Anais...** Belo Horizonte: SBC, 2010.

VOORHEES, E. M. Query expansion using lexical-semantic relations. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 17., New York, NY, USA. **Proceedings...** Springer-Verlag New York: Inc., 1994. p.61–69. (SIGIR '94).

WILKINSON, R.; HINGSTON, P. Using the cosine measure in a neural network for document retrieval. In: SIGIR - SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL FORUM. **Anais...** ACM, 1991.

WILLIAM, M. **The Research Methods Knowledge Base, 2nd Edition**. Disponível em: <http://www.socialresearchmethods.net/kb/>. Acesso em: maio 2011.

XU, J.; CROFT, W. B. Query expansion using local and global document analysis. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 19., New York, NY, USA. **Proceedings...** ACM, 1996. p.4–11. (SIGIR '96).

XU, J.; CROFT, W. B. Improving the effectiveness of information retrieval with local context analysis. **ACM Trans. Inf. Syst.**, New York, NY, USA, v.18, p.79–112, January 2000.

ZHANG, J.; DENG, B.; LI, X. Concept Based Query Expansion Using WordNet. In: INTERNATIONAL E-CONFERENCE ON ADVANCED SCIENCE AND TECHNOLOGY, 2009., Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2009. p.52–55. (AST '09).