ATLASOM: PROCESSAMENTO DE TEXTO PARA A GERÊNCIA DE UMA COLABORAÇÃO CIENTÍFICA DE GRANDE PORTE

Felipe Fink Grael

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: José Manoel de Seixas

Rio de Janeiro

Junho de 2013

ATLASOM: PROCESSAMENTO DE TEXTO PARA A GERÊNCIA DE UMA COLABORAÇÃO CIENTÍFICA DE GRANDE PORTE

Felipe Fink Grael

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. José Manoel de Seixas, D.Sc.

Prof. Luiz Pereira Calôba, D.Sc.

Prof. Francisco de Assis Tenório de Carvalho, D.Sc.

Profa. Érica Polycarpo, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

JUNHO DE 2013

Grael, Felipe Fink

ATLASom: Processamento de Texto para a Gerência de uma Colaboração Científica de Grande Porte/Felipe Fink Grael. – Rio de Janeiro: UFRJ/COPPE, 2013.

XII, 47 p.: il.; 29, 7cm.

Orientador: José Manoel de Seixas

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2013.

Referências Bibliográficas: p. 42 – 47.

1. ATLAS. 2. SOM. 3. Mineração de Dados em Texto. I. Seixas, José Manoel de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

Agradeço primeiramente a minha família. Aos meus pais, Christiane e Oscar Grael, pelo amor e apoio incondicionais. A minha tia Luciane, pelo incentivo e carinho desde sempre. Seus ensinamentos e atenção foram fundamentais para mim.

Ao meu orientador, Prof. José Manoel de Seixas, com quem aprendi muito. A Carmen Maidantchik, por toda a atenção e oportunidades. Ao longo desses dez anos de trabalho, vocês foram fundamentais para minha formação acadêmica e profissional.

A Fernando Ferreira, Andressa Sivolella, Laura Moraes, Luiz Évora e Carlo Fragni, meus sócios e, sobretudo, grande amigos. Agradeço a todo o incentivo, compreensão e ajuda que sempre me deram, e pelo companheirismo nesse caminho empreendedor. É um privilégio a oportunidade de trabalhar e conviver com pessoas brilhantes como vocês.

A toda colaboração com o ATLAS, em especial a Kathy Pommès. Agradeço a oportunidade de trabalhar nesse ambiente fantástico e singular, bem como o apoio a mim e ao grupo de trabalho do qual faço parte.

A todos meus colegas do Laboratório de Processamento de Sinais (LPS) e da COPPE/UFRJ. O prazer de trabalhar, aprender e conviver com vocês são um grande incentivo para continuar por esse caminho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ATLASOM: PROCESSAMENTO DE TEXTO PARA A GERÊNCIA DE UMA COLABORAÇÃO CIENTÍFICA DE GRANDE PORTE

Felipe Fink Grael

Junho/2013

Orientador: José Manoel de Seixas

Programa: Engenharia Elétrica

Estima-se que 53% de todas as informações geradas dentro de uma empresa sejam semi ou não estruturadas. No caso do detector de partículas ATLAS, que é operado por uma colaboração internacional de grande porte, envolvendo 38 países, as trocas de informação se dão por e-mail, apresentações em reuniões, páginas web escritas de forma colaborativa e artigos. Dado a alta rotatividade de colaboradores e a grande mobilidade entre áreas, é muito importante para a gerência saber a área de trabalho de cada pessoa. Neste trabalho é estabelecida uma metodologia para recuperar os documentos que são gerados diariamente pelos colaboradores, e agrupá-los de acordo com a área de trabalho. Com isso, possibilita-se relacionar os autores e outras pessoas com sua área de colaboração.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ATLASOM: TEXT PROCESSING FOR THE MANAGEMENT OF A LARGE SCALE SCIENTIFIC COLLABORATION

Felipe Fink Grael

June/2013

Advisor: José Manoel de Seixas

Department: Electrical Engineering

Around 53% of all information generated in business organizations are semi-structured or unstructured. Concerning the ATLAS particle detector, which is run by a large scale international collaboration spanning 38 countries, information are exchanged mainly by e-mail, presentations on meetings, papers and collaboratively written web pages. Due to the turnover rate, and the high mobility the collaborators have among the different areas, it is of big value for the management to keep track of the relation between areas and people. This work establishes a process for retrieving documents that are generated in the everyday work of the collaborators and clustering them according to the activity area.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ATLAS	*A Toroidal LHC Apparatus*, p. 2
BI	*Business Intelligence*, p. 15
CDS	*CERN Document Server*, p. 11
CERN	Organização Europeia para Pesquisa Nuclear, p. 2
EM	*Expectation Maximization*, p. 23
GTM	*Generative Topographic Map*, p. 23
LAr	*Liquid Argon*, p. 7
LHC	*Large Hadron Collider*, p. 2
LSI	Indexação por Semântica Latente, p. 20
PCA	Análise de Componentes Principais, p. 19
SOM	*Self-Organizing Maps*, p. 21
SVD	Decomposição em Valores Singulares, p. 19
TF-IDF	*Term Frequency-Inverse Document Frequency*, p. 18
TileCal	Calorímetro de Telhas, p. 7
VSM	*Vector Space Model*, p. 17

Capítulo 1

Introdução

A produção e troca de informações é parte vital de qualquer tipo de organização. Portanto, informações de diversos tipos são continuamente produzidas e armazenadas em diferentes repositórios. Por exemplo, registros de compra e venda podem ser armazenados em bancos de dados, enquanto decisões estratégicas tomadas em reuniões podem ser armazenadas em documentos textuais, apresentações ou enviadas por e-mail. Como consequência, uma boa parte do conhecimento, ou seja, aquilo que pode-se aprender a partir das informações dessas organizações ficam espalhados por esses documentos.

É cada vez mais comum o uso de sistemas de gerenciamento de documentos, a fim de disponibilizar essas informações para os envolvidos. Frequentemente esses sistemas fornecem busca sobre os metadados (por exemplo título do documento, tipo, autor, data de criação), e às vezes até sobre o conteúdo. No entanto, por mais fácil que seja recuperar um arquivo, o conhecimento em si não é capturado por essas ferramentas. Para responder questões gerenciais do tipo "quem está trabalhando em um determinado setor", ou "quem possui conhecimentos em determinada área", as ferramentas não deixam outra opção senão ler esses documentos para realizar essas análises. Quanto maior as organizações, mais importantes são esses tipos de análise, embora mais difícil seja para uma pessoa acompanhar o crescente volume de informações.

A mineração de dados em texto é uma área da Inteligência Computacional que tem recebido muita atenção nos últimos anos. As técnicas especificamente desenvolvidas para texto podem ser combinadas de inúmeras formas com os métodos clássi-

cos de aprendizagem de máquina, tornando extremamente instigantes as pesquisas voltadas a extrair conhecimento de um grande volume de dados não estruturados.

1.1 Motivação

O acelerador de partículas LHC (*Large Hadron Collider*) é atualmente o maior e mais potente já construído, e opera no CERN (Organização Europeia para Pesquisa Nuclear). Em torno dos seus quatro pontos de colisão são instalados detectores de partículas, com o objetivo de estudar diferentes aspectos da física de altas energias. Esses detectores foram construídos e são agora operados por colaborações envolvendo centenas de institutos em diversos países.

O presente projeto está inserido no contexto da colaboração internacional do ATLAS (*A Toroidal Lhc Apparatus*), o maior dos detectores de partículas do LHC. A colaboração é composta por mais de 4000 membros, provenientes de 177 institutos espalhados por 38 países. O ATLAS, que é o maior dentre os detectores que operam no LHC, possui formato cilíndrico, medindo cerca de 45m de comprimento e 25m de diâmetro. Esse grande aparelho é dividido em 4 sub-detectores, cada um associado a uma área da colaboração. Existem ainda outras áreas na colaboração que são relacionadas a partes mais operacionais, como a Coordenação Técnica e Computação, e também áreas voltadas a análise dos dados à luz das teorias físicas sendo estudadas.

A troca de informações entre os colaboradores se dá através de diferentes formas: documentos apresentados em reuniões, páginas web escritas de forma colaborativa através da ferramenta tipo *wiki*, artigos publicados como notas internas ou em revistas e conferências científicas, dentre outros. Para cada uma dessas formas existem sistemas que concentram as informações, no intuito de facilitar uma posterior consulta, e mantê-la disponível para todos os interessados.

O grande volume de informações, no entanto, dificulta o processo de entrada de um novo colaborador em uma área no qual ele não tenha familiaridade, visto que o conhecimento está fragmentado por diversos documentos. Além disso, dada a alta rotatividade de pessoas por entre os setores da colaboração, é importante, do ponto de vista gerencial, saber a alocação dos membros, para garantir que não haverá áreas

pouco supridas.

1.2 Objetivos

Este trabalho aplica técnicas de mineração de dados em texto e inteligência computacional com o intuito de agrupar os documentos da colaboração por área. Ao agrupar por área, é possível, posteriormente, inferir qual o setor de atuação do autor e demais membros relacionados ao documento.

Dois importantes instrumentos da colaboração são as páginas *wiki* armazenadas no sistema TWiki e os artigos, teses e notas internas que são publicadas no *CERN Document Server* (CDS). Ambos são sistemas acessíveis via Web e, portanto, fácil de serem acessados.

O objetivo do presente trabalho é estabelecer um processo capaz de agrupar os documentos proveniente dessas fontes. Para isso os documentos são transformados em representações vetoriais para serem agrupados usando uma combinação de mapas auto-organizáveis (SOM, da sigla em inglês para *Self Organizing Maps*), e o algoritmo clássico de agrupamento *k-means*.

1.3 Organização do documento

O Capítulo 2 apresenta o ambiente do CERN e a colaboração ATLAS, explicando os documentos gerado pelos colaboradores. O Capítulo 3, por sua vez, apresenta as técnicas de processamento de texto e inteligência computacional que serão empregadas nesse trabalho para realizar o agrupamento por área. Em seguida, o Capítulo 4 apresenta a forma que essas técnicas foram utilizadas, e como foram realizados os experimentos. Os resultados experimentais são então apresentados no Capítulo 5. Por fim, as conclusões obtidas e propostas de trabalhos futuros são apresentadas no Capítulo 6.

Capítulo 2

A Colaboração ATLAS do CERN

A mineração de dados em textos é uma área da Inteligência Computacional que tem recebido muita atenção nos últimos anos. As técnicas utilizadas são particularmente interessantes para o âmbito de gerenciamento baseado em conhecimento. Organizações e corporações procuram se tornar cada vez mais flexíveis frente ao cenário globalizado e competitivo [1]. Como grande parte do volume de informações encontra-se desestruturado, principalmente em documentos de texto, a mineração de dados torna-se uma ferramenta fundamental.

O cenário que motivou o presente trabalho tem como contexto o experimento ATLAS, localizado no CERN. O tamanho da colaboração é um desafio por si só. Manter a relação das áreas de atuação dos seus mais de 3.000 membros é de grande importância para a colaboração. Por possuir uma hierarquia horizontal, onde diferentes grupos atuam em paralelo e são geridos em uma estrutura matricial [2], torna-se uma missão complicada compreender qual a contribuição de cada membro do experimento. Essa tarefa é ainda dificultada pelo caráter fluído da colaboração, onde há uma grande rotatividade das pessoas envolvidas.

Atualmente, esse processo de mapeamento das atividades da colaboração é falho, pois é feito no momento do cadastramento do contrato do membro. Mesmo que seja permitido ao líder do grupo editar essa informação, observa-se que isso raramente acontece, tornando a atual base de dados uma fonte não confiável em relação à área de atuação de cada colaborador.

O CERN (*Organisation européenne pour la recherche nucléaire*) é o maior laboratório de física de partículas do mundo e nele se encontra o mais complexo aparato

científico já construído pelo homem, o LHC (*Large Hadron Collider*) [3]. Neste capítulo será apresentado o experimento ATLAS, instalado neste acelerador, com enfoque em seus subsistemas, sua estrutura gerencial e nas ferramentas que, além de darem suporte à esta gerência, serviram como fonte de dados para este trabalho.

2.1 Organização Europeia para Pesquisa Nuclear

O CERN localiza-se na fronteira entre a Suíça e a França, próximo a cidade de Genebra. Foi fundado com o objetivo de criar um laboratório europeu para pesquisa de física nuclear, em que seus países membros pudessem dividir as despesas de seus aparatos e instalações [4]. Atualmente, o CERN é dirigido por 20 Estados membros europeus. Ainda, diversos países não-membros colaboram de diferentes maneiras, inclusive o Brasil. No total, existem cerca de 8.000 cientistas provenientes de 608 universidades, representando 113 nacionalidades [5].

Os benefícios das pesquisas realizadas no CERN vão além da comunidade de física de partículas. As tecnologias nele desenvolvidas são empregadas em diversas áreas. A criação da *World Wide Web*, equipamentos de imagens medicinais, painéis solares mais eficientes e a computação em GRID são apenas alguns dos exemplos.

2.2 O Projeto LHC

O LHC, o maior e mais poderoso acelerador de partículas do mundo, é o mais novo aparato do complexo de aceleradores do CERN. Ele está instalado em um túnel que possui uma circunferência de 27 km de extensão e está posicionado a 150 metros abaixo da superfície. Antes de serem injetados com uma energia de 450 GeV no anel do LHC, prótons são acelerados formando feixes. Esses feixes são acelerados até que a sua a energia aumente cerca de 15 vezes, adquirindo 7.000 GeV. Como o objetivo do projeto demanda uma elevada taxa de tomada de dados, o LHC opera a 40 MHz. Operando na luminosidade máxima projetada para o colisionador, a taxa de eventos pode alcançar 1 GHz [3].

Seu formato permite que vários experimentos sejam realizados nos diversos pontos de colisão ao longo de sua circunferência, que possibilitam uma visão detalhada das colisões ocorridas, indo desde características energéticas às imagens com alta

resolução da trajetória das partículas resultantes. De posse das informações proporcionadas pelos detectores, podemos caracterizar cada sub-partícula criada após a colisão.

Quatro grandes experimentos estão localizados ao redor da circunferência do LHC, como mostrado na Figura 2.1: ATLAS [6, 7], CMS [8], LHCb [9] e ALICE [10].

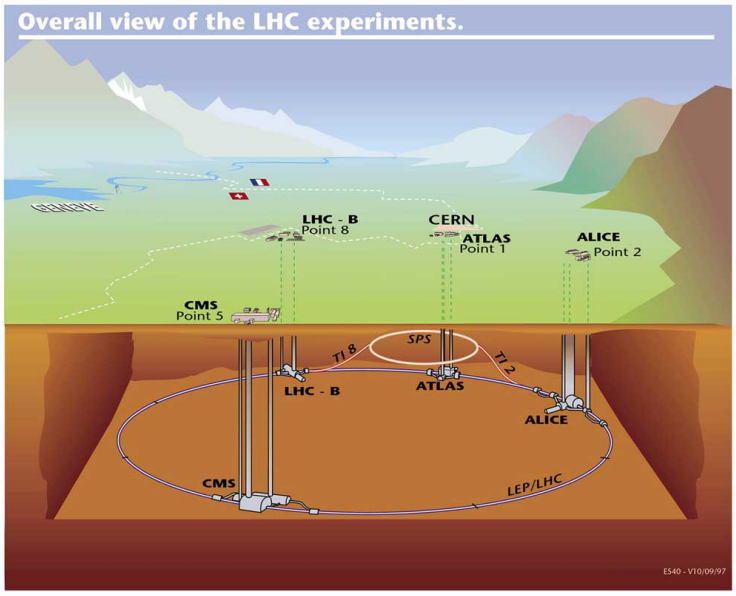


Figura 2.1: Diagrama esquemático do túnel do LHC (Extraído de [5]).

O ATLAS e o CMS são experimentos de proposta geral, otimizados para estudar possíveis processos decorrentes de uma física ainda desconhecida, que podem ocorrer a energias na faixa dos TeV. Os outros dois foram projetados para estudar fenômenos específicos: o LHCb realiza medições precisas de decaimentos dos mésons B e D, e o ALICE estuda colisões de íons pesados.

2.3 O Experimento ATLAS

O detector ATLAS tem aproximadamente 45 m de comprimento, mais de 25 m de altura e pesa em torno de 7.000 T. Ao todo, 3.000 pesquisadores de 38 países participam da colaboração, representando 174 instituições diferentes [11].

O detector é composto por subsistemas distintos, com características e objetivos específicos. A Figura 2.2 apresenta o detector ATLAS e os sistemas que o compõe.

Resumidamente, pode-se dividir o experimento ATLAS em quatro detectores principais [13]:

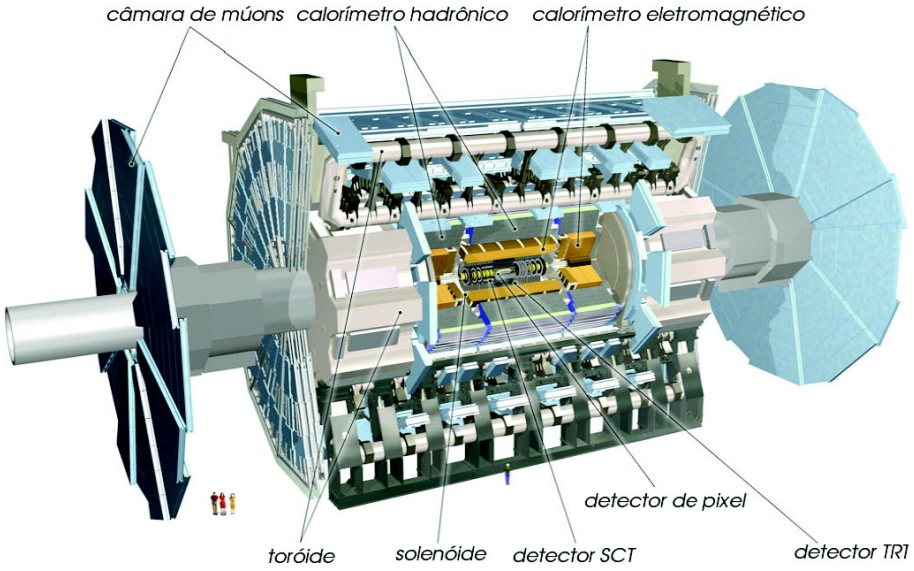


Figura 2.2: Diagrama ilustrativo do detector ATLAS (adaptado de [12]).

Detector de Traços composto por uma infinidade de sensores altamente segmentados feitos de silício, determina com muita precisão a trajetória de partículas carregadas. Estas trajetórias são desviadas devido ao campo magnético presente no ATLAS. É possível dessa forma estimar o momento e a carga da partícula passante [14]. Os detectores SCT, TRT e de *pixel* são parte deste subsistema. Geralmente esse conjunto de sub-detectores é chamado de *Inner Detector*.

Calorímetro Eletromagnético determina a energia e o perfil de deposição de energia de partículas eletromagnéticas, como, por exemplo, elétrons e fótons. Está dividido em quatro camadas, cada uma com granularidade distinta [15]. Devido ao fato de ser formado por condutores submersos em argônio líquido, é comumente chamado de *Liquid Argon*, ou LAr.

Calorímetro Hadrônico determina a energia e o perfil de deposição de energia de partículas hadrônicas, como por exemplo os prótons. Está dividido em 3 camadas com granularidades distintas [16]. Por empregar telhas cintilantes, esse calorímetro costuma ser referido por Calorímetro de Telhas ou *TileCal*.

Espectrômetro de múons O Espectrômetro é composto pela câmaras de múon [17], onde um campo magnético (produzido pelos toróides) é utilizado para curvar a trajetória dos múons. Este procedimento aliado a um poderoso

sistema de detecção de traço específico, permite a medição do momento da partícula.

2.3.1 Modelo Gerencial do ATLAS

A organização gerencial do ATLAS é respaldada em quatro princípios: democracia, mínima organização formal, mandatos limitados e separação entre os poderes de criação de políticas e executivo. Sua estrutura reflete os três eixos que moldam a colaboração: a física do experimento, o detector como aparato experimental e a cultura colaborativa [18].

O mapa apresentado na Figura 2.3 apresenta as trinta e oito diferentes nações em que existem institutos participando da colaboração ATLAS. São mais de 3.000 cientistas (aproximadamente 1.000 alunos de doutorado) trabalhando no projeto e produzindo os artigos que serão publicados [2]. O ATLAS representa, desta maneira, uma colaboração global, trabalhando de maneira cotidiana e contínua.

Figura 2.3: Países que participam da colaboração ATLAS.

Porta-voz

A estrutura organizacional do ATLAS, em março de 2013, está representada no esquema da Figura 2.4. O líder da colaboração é o *spokesperson*, o "porta-voz" em livre tradução. Seu papel é delegar responsabilidades, guiar todos os aspectos do projeto e quando necessário, tomar decisões, com a ajuda do *collaboration board*. O *technical coordinator* é o responsável por fazer todos os detectores do experimento estarem integrados e em operação harmônica. Já o *Resources Coordinator* é responsável por planejar os recursos disponíveis e garantir que a demanda do ATLAS seja condizente com a disponibilidade das diferentes nações envolvidas [19].

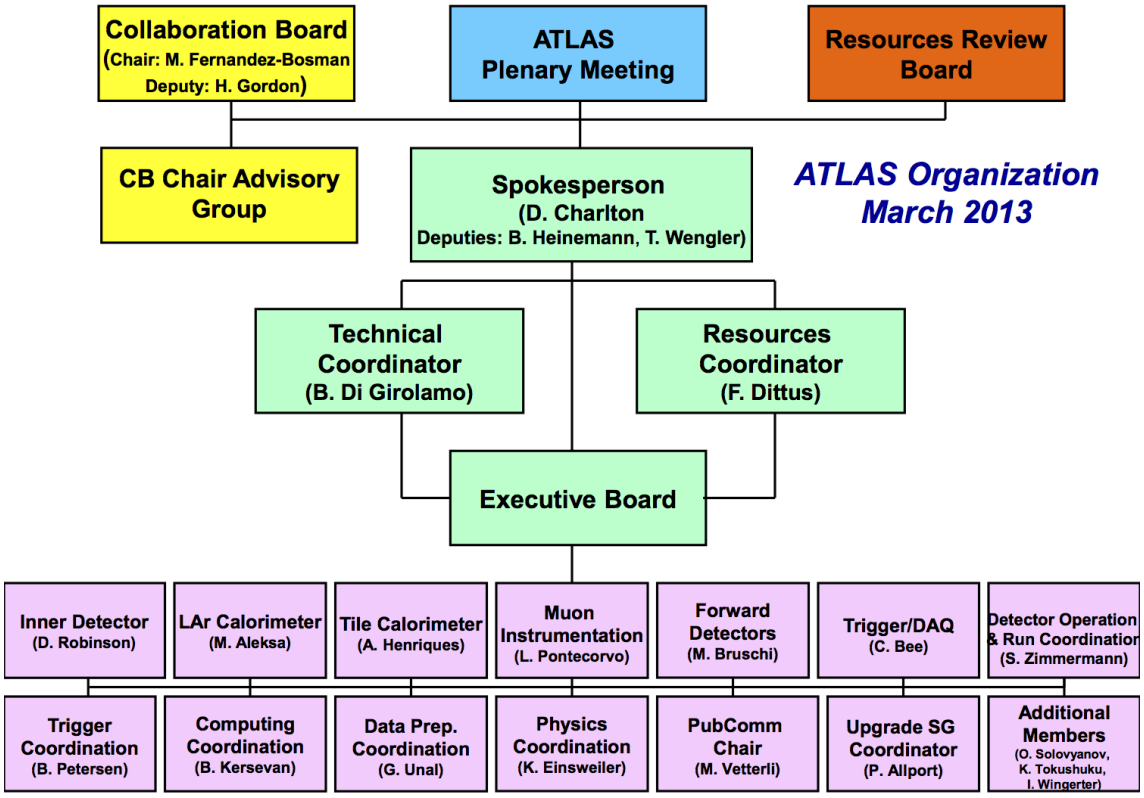


Figura 2.4: Organização da Colaboração ATLAS em 2013.

Plenary Meeting

A Reunião Plenária, ou *Plenary Meeting*, é o fórum aberto para participação de todos da colaboração. Todas as decisões referentes a resultados e objetivos físicos; operação, mudanças e atualização do detector; e questões organizacionais são discutidas no plenário e, se necessário, nos grupos de trabalho subordinados.

Collaboration Board

O *Collaboration Board* é o conselho responsável por tomadas de decisões sobre as políticas do ATLAS. Algumas das tarefas do conselho são:

- decisões sobre mudanças no detector e atualizações;
- políticas internas;
- recursos financeiros e humanos;
- eleições;
- organização gerencial do ATLAS;
- organização de membros do ATLAS;

As reuniões do conselho ocorrem três vezes ao ano.

Executive Board

O *Executive Board* direciona a execução do projeto ATLAS, de acordo com as políticas definidas pelo *Collaboration Board*. Este conselho executivo aproxima os coordenadores responsáveis pela operação do ATLAS e a exploração da física, com a visão global do desempenho do experimento. Algumas das atividades desse conselho são descritas a seguir:

- estabelecer o programa de Física;
- monitorar a operação e manutenção dos sub-detectores;
- monitorar a manipulação dos dados e computação;
- gerenciar recursos humanos;
- coordenação das atividades entre sub-detectores;
- coordenação de mudanças de *software* e *hardware*

A composição do *Executive Board* é adaptada conforme as necessidades momentâneas do experimento. Essa composição, bem como a descrição das tarefas de seus membros é aprovada pelo *Collaboration Board*.

As reuniões deste conselho ocorrem mensalmente. Há ainda a reunião da gerência do ATLAS, em que, semanalmente, o *spokesperson*, os *deputies*, o *resource coordinator* e o *technical coordinator* participam.

2.3.2 Ferramentas da Colaboração

O CERN possui um conjunto de ferramentas para apoiar a organização dos dados e a troca de informação entre os membros [20]:

Indico Ferramenta usada para agendar eventos, tais como reuniões, conferências ou palestras. Todos os materiais utilizados no evento, tais como apresentações, minutas de reuniões ou documentos de referência são armazenados no sistema. Também suporta eventos públicos ou privados. [21, 22].

CDS (*CERN Document Server*) É o banco oficial principal para todos os documentos do CERN. Todos os artigos publicados, rascunhos e notas internas são registrados, permitindo aos colaboradores consultar e fazer comentários. Tem suporte tanto para documentos públicos quanto para restritos [23, 24].

CERN TWiki É uma ferramenta para páginas web colaborativas. A princípio qualquer colaborador pode atualizar o conteúdo de uma página através da própria interface web. Como os sistemas anteriores, também permite que sejam criadas não só páginas públicas mas também restritas a um grupo específico [25]. O ATLAS começou a usar o sistema TWiki em 2005 e atualmente a maior parte das páginas referentes ao experimento estão no formato *wiki* [26].

Os dois principais tipos de páginas TWiki usadas no ATLAS são:

`ATLAS` Páginas públicas contendo informações técnicas sobre o ATLAS;

`AtlasProtected` páginas restritas com as informações necessárias para preparação dos dados físicos, manipulação de algoritmos e publicação interna dos resultados.

E-mails Todos os colaboradores do CERN possuem uma conta de e-mail e pode fazer parte ou criar grupos eletrônicos (*e-groups*). É a forma clássica de se comunicar entre pessoas, e uma das mais usadas para se comunicar com grandes grupos.

Sistema Glance Mecanismo genérico para acesso a qualquer banco de dados. Funciona como uma camada intermediária, isolando o usuário das particularidades de cada banco de dados. Com ele é possível recuperar, inserir e atualizar registros estruturados em bancos de dados. [27].

2.3.3 Documentos da Colaboração

A colaboração ATLAS produz um elevado número de documentos espalhados pelas ferramentas usadas para compartilhamento das informações. Com exceção do correio eletrônico, todos os demais sistemas são acessíveis via Web, tornando o acesso fácil. Algumas das estatísticas de utilização são fornecidas a seguir.

Sistema Indico

O sistema Indico é intensamente usado por todo o CERN. Só para a colaboração ATLAS existem 100.000 eventos atualmente cadastrados, com mais de 440.000 documentos inseridos. A Figura 2.5 apresenta a quantidade de documentos submetidos ao sistema desde 2000 até junho de 2013.

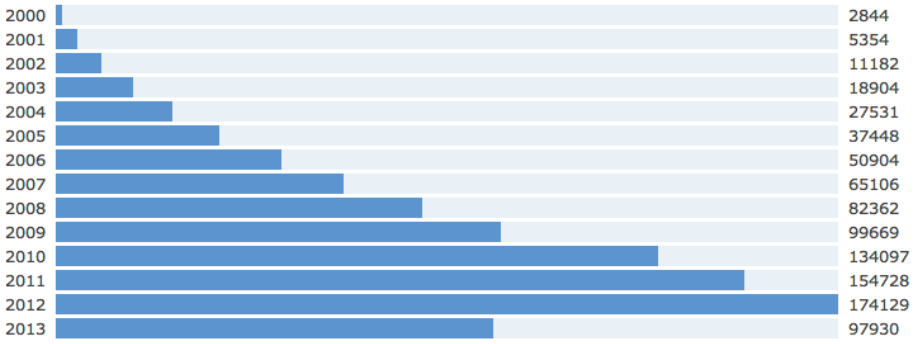


Figura 2.5: Número de documentos inseridos no Indico de 2001 até junho de 2013. Retirado de [21].

Plataforma CDS

Com a operação do LHC, observou-se um aumento significativo no número de artigos, relatórios e teses referentes aos experimentos do acelerador. Para o ATLAS, em 2010 foram publicados 32 artigos. Em 2011, esse número foi duplicado, passando para 64 [20]. Desde janeiro de 2012, foram aprovados 155 artigos. Para o mesmo período, foram publicadas 225 notas de conferência. Um aumento de 40% em relação a 2011.

A Figura 2.6 mostra a estatística de todos os documentos inseridos na plataforma CDS por todo o CERN desde 2011.

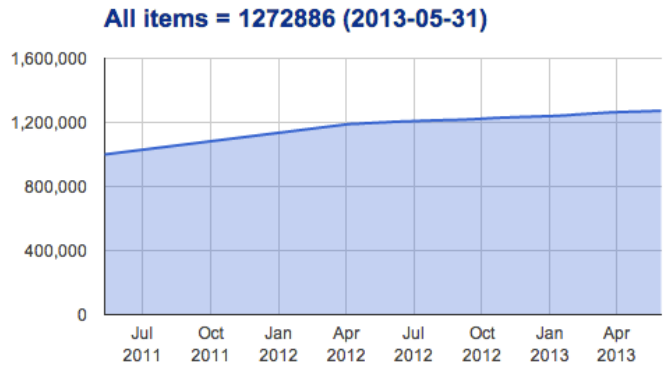


Figura 2.6: Número de documentos criados na plataforma CDS desde 2006. Retirado de [28].

Plataforma TWiki

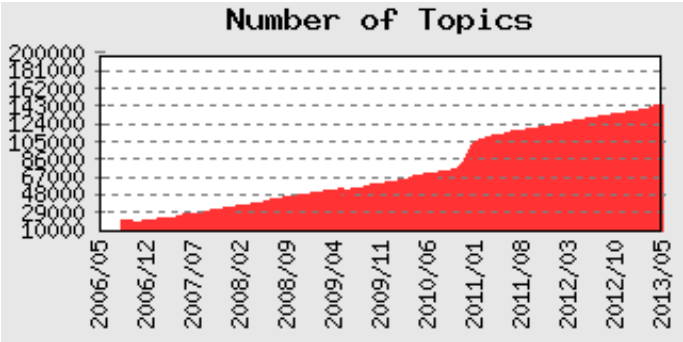


Figura 2.7: Número de documentos criados na plataforma TWiki desde 2006. Retirado de [29].

A taxa de criação de novos documentos para a colaboração ATLAS na plataforma TWiki é de 150 novas ocorrências por mês [26] e há uma média de 10.000 atualizações para o mesmo período. Só em 2013, mais de 1.000 novas páginas foram criadas. A

Figura 2.7 apresenta a evolução de criação de páginas novas, desde 2006, para todas as colaborações do CERN. Nota-se um aumento no volume de páginas após o início da operação em 2010.

Capítulo 3

Inteligência Computacional Aplicada a Textos

As informações que são geradas e armazenadas diariamente nas organizações podem ser classificadas em três grandes tipos: estruturadas, não-estruturadas, ou semi-estruturadas. Por dados estruturados entende-se aqueles que obedecem uma modelagem previamente estabelecida, tais como os que são armazenados em bancos de dados relacionais. Os não-estruturados, por sua vez, são aqueles que são armazenados sem qualquer estrutura, tais como textos lives, imagens, etc. Em uma categoria intermediária ficam os semi-estruturados, que são formados predominantemente por informações não-estruturadas, mas possui algum tipo de marcação que determina sua estrutura. Um artigo científico, por exemplo, pode ser considerado um documento semi-estruturado pois, ao longo do conteúdo, estão claramente marcado diferentes elementos estruturais, como o título, o resumo (*abstract*), os autores, e o corpo em si do texto.

Uma pesquisa realizada em 2006 com profissionais da área de Inteligência Empresarial (BI, da sigla em inglês para *Business Intelligence*) estima que 53% de todas as informações dentro de uma empresa são não-estruturadas ou semi-estruturadas [30]. Outras pesquisas, que usam como base profissionais de outros contextos, chegam a estimar que esses tipos de informação correspondam a 85%. Isso mostra a grande importância e potencial de aplicação das técnicas para mineração de texto para organizações em geral.

Mineração em dados textuais encontra aplicações em diferentes áreas, tais como

ordenar documentos por importância [31], categorização automática de notícias [32], predição de cancelamento de serviço em centrais de relacionamentos com clientes [33], e até identificação de autores de e-mails em análises forenses [34].

De acordo com a aplicação, diferentes aspectos do texto são levados em consideração. Para classificação de documentos por similaridade, por exemplo, o trabalho WEBSOM [35] aplica mapas auto-organizáveis a textos não-estruturados. Cada documento é representado por um vetor em um espaço onde suas dimensões são determinadas unicamente a partir das palavras que ocorrem no conjunto de documentos de treino.

Este capítulo descreve as técnicas usadas no presente trabalho para realizar mineração de dados no contexto da colaboração ATLAS, tanto no que se refere à representação de dados textuais de forma vetorial quanto às técnicas de inteligência computacional relacionadas.

3.1 Pré-processamento de texto

O primeiro passo da abordagem utilizada neste trabalho para processar um documento textual é dividí-lo em frases e, em seguida, em palavras. A divisão do texto completo em frases é feita utilizando o algoritmo Punkt [36], por ser robusto a abreviaturas. Este algoritmo foi testado com sucesso em textos em diversos idiomas. No entanto, pode-se assumir que os textos em estudo estão em inglês, uma vez que esse é o idioma predominante para comunicação dentro da colaboração. A divisão de frase em palavras é feita com o algoritmo utilizado no projeto Penn Treebank [37], que além de separar as palavras por espaço em branco, separa contrações como "*they'll*" em "*they*" e "*'ll*", o que é adequado para textos em inglês.

Uma vez tendo os documentos representados como lista de palavras, faz-se uma filtragem inicial para remover uma série de palavras comuns e que não contribuem para o significado do texto. Essas palavras são denominadas "palavras vazias" (*stopwords*), e uma lista predefinida de palavras em inglês foi utilizada.

Em seguida, as palavras passam pelo processo de lematização, no qual é extraído somente o radical das palavras. Para isso foi utilizado o lematizador de Porter [38]. Com esse algoritmo, as palavras "*lies*" e "*lying*" são revertidas para o

radical "*lie*", e podem ser tratadas como tendo o mesmo valor semântico para a análise dos documentos. Algumas palavras são convertidas para outras formas que não são exatamente seu radical. A palavra "ATLAS", por exemplo, é convertida para "ATLA". Isso, no entanto, não prejudica o estudo, visto que todas as ocorrências serão consistentemente convertidas da mesma forma.

A Figura 3.1 exemplifica as etapas descritas acima, utilizando um documento hipotético de duas frases. Antes da escolha do modelo vetorial, detalhado na Seção 3.2 ainda são removidas as pontuações, números ou partículas que não formam palavras, tais como "*'s*".

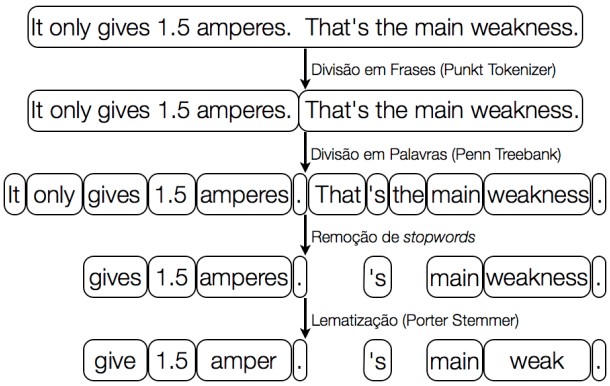


Figura 3.1: Exemplo de pré-processamento sobre um documento.

É possível encontrar diversas implementações desses algoritmos. No presente trabalho, faz-se o uso da biblioteca NLTK [39], para a linguagem de programação Python. Essa biblioteca inclui ainda a lista de *stopwords* utilizada, composta por 127 palavras.

3.2 Representação vetorial

Para que documentos textuais possam ser comparados entre si ou com um determinado padrão, é necessário representá-los algebricamente. A forma mais utilizada é o modelo de espaço vetorial (VSM, na sigla em inglês para *Vector Space Model*) [40]. Nessa representação vetorial, cada dimensão desse vetor é um peso correspondente à ocorrência de um determinado termo no documento.

Há diversas formas de determinar o peso para a representação vetorial. A mais comum é usar um valor proporcional à frequência do termo no documento, mas inversamente proporcional ao número de documentos no qual está presente. Essa

forma é denominada TF-IDF (da sigla em inglês para *term frequency–inverse document frequency*) [41].

O fator proporcional à frequência da palavra, o *tf*, faz com que o termo seja maior à medida em que o número de ocorrências dessa palavra aumente. Uma das formas de calcular esse termo é mostrado abaixo, dado que $n_{i,j}$ é o número de ocorrências da palavra i no documento j, e K o total de palavras que compõem o espaço vetorial

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{K} n_{k,j}}$$

Já o fator *idf* é responsável por diminuir o peso de uma palavra, caso ela ocorra em muitos documentos, visto que nesse caso ela agrega pouca informação ao texto. Vale notar que o *idf* é um parâmetro global, ou seja, seu valor é igual para todos os documentos, variando somente com a palavra. Uma das formas de calculá-lo é como abaixo, dado que d é o conjunto de termos de um determinado documento pertencente a D, que é o conjunto de todos documentos, e t_i denota o termo i do modelo vetorial.

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

Para montar o espaço vetorial deve-se primeiramente decidir a lista de palavras que serão levadas em consideração. Para isso, deve-se observar a ocorrência de cada palavra no conjunto de documentos disponível para treinamento, já pré-processados.

O vetor final terá, então, tantas dimensões quanto palavras escolhidas, e o valor desse vetor será o produto $\mathrm{tf}_{i,j} \times \mathrm{idf}_i$.

No caso de documentos semi-estruturados, para fazer uso da informação que a estruturação do documento provê, deve-se realizar o estudo de seleção de palavras e cálculo dos valores TF-IDF para cada um dos atributos, separadamente [42]. A representação vetorial final do documento será um vetor aumentado, correspondente a concatenação dos vetores dos atributos.

3.3 Redução de Dimensionalidade

Como discutido na seção anterior, a representação vetorial tem tantas dimensões quanto palavras que sejam escolhidas para representar um conjunto de documentos.

Dependendo do tipo de documento sendo analisados, o número de palavras distintas pode chegar facilmente na ordem de centena de milhares. Um vetor nessa ordem de dimensões, no entanto, é prejudicial às análises, tanto por aumentar o esforço computacional para processamento das informações quanto pelo efeito da "maldição da dimensionalidade" [43]. Por isso, faz-se necessário aplicar algum método de redução de dimensionalidade.

Uma das formas de reduzir o número de dimensões é simplesmente filtrar as palavras escolhendo somente as que são usadas em um número de documentos dentro de um intervalo máximo e mínimo. A razão para isso seria que uma palavra frequente demais não agregam informação no conjunto de documentos analisados, enquanto as raras demais não sejam representativas para a coleção. No entanto, determinar os limites desse intervalo de forma razoável é uma questão em aberto. Outros dois métodos que mostram resultados promissores são a Análise de Componentes Principais e a Projeção Aleatória, descrito nas subseções a seguir.

3.3.1 Análise de Componentes Principais

A Análise de Componentes Principais (PCA, da sigla em inglês para *Principal Component Analysis*) é uma abordagem clássica para compactação de sinais e extração de características [44]. A análise parte do princípio que há correlação entre os componentes dos dados originais e, portanto, redundância de informação. O objetivo então é encontrar uma transformação linear tal que, no novo espaço, a redundância introduzida pela correlação seja eliminada. Ao mesmo tempo, prevê-se que os componentes no novo espaço são reordenados em ordem decrescente de variância, ou seja, o primeiro componente será o de maior variância do sinal.

A forma fechada de encontrar uma projeção que atenda o critério de descorrelacionamento é através da base formada pelos autovetores da matriz de correlação C_x, onde x são os vetores do sinal original centrados, ou seja, que tiveram a média removida. Esses autovetores serão então os componentes principais, e os autovalores correspondentes darão a variância, ou seja, a energia do sinal correspondente ao componente. Tipicamente usa-se o método de decomposição em valores singulares (SVD, da sigla em inglês para *Singular-value decomposition*).

A compactação do sinal pode ser feita selecionando-se somente os componen-

tes com maiores autovalores. Analisando-se a soma cumulativa dos autovalores já ordenados, pode-se determinar o número de componentes baseado na quantidade de energia total do sinal original. Em casos práticos, a maior parte da energia se deposita sobre um conjunto pequeno de componentes principais.

Um método correlacionado frequentemente utilizado é o LSI (da sigla em inglês para *Latent Semantic Indexing*) [45]. Esse método consiste em realizar a SVD diretamente sobre a matriz de correlação dos vetores que representam os documentos por TF-IDF. A diferença entre PCA e LSI é que o primeiro remove a média dos componentes. Ao manter a média na análise, perde-se a correspondência entre os autovalores e a energia do sinal original, que é uma informação útil na análise do processo. No entanto, ao centralizar os componentes, perde-se a esparsidade característica da matriz TF-IDF.

Devido à vantagem da interpretação matemática dos autovalores, escolheu-se pela PCA para os estudos no presente trabalho.

3.3.2 Projeção Aleatória

A Projeção Aleatória (RP, da sigla em inglês para *Random Projection*) é uma forma de baixo esforço computacional para redução de dimensionalidade [46]. Consiste em mapear o espaço original $\mathbb{R}^N$ em um espaço reduzido $\mathbb{R}^d$, onde $d \ll N$ através de vetores cujos valores são amostras de uma distribuição com média zero. A matriz de projeção deve ser montada como um conjunto de d vetores de norma unitária.

Apesar de ser contra-intuitivo, mostra-se que uma projeção desse tipo, dado que a dimensão d seja ainda suficientemente grande, mantém a estrutura de distâncias do espaço original [46]. Dado dois vetores $\mathbf{x_1}$ e $\mathbf{x_2}$ no espaço original, uma forma de medir a distância entre eles é a similaridade de cossenos, que pode ser calculada através do produto interno desses vetores. Sendo $\mathbf{y_1}$ e $\mathbf{y_2}$ o resultado da projeção desses dois vetores por uma matriz R, a similaridade de cossenos entre eles pode ser estabelecida como:

$$\mathbf{y_1}\mathbf{y_2}^T = \mathbf{x_1}RR^T\mathbf{x_2}^T$$

Se todos os vetores que compõem a matriz R fossem ortogonais, o resultado de RR^T seria a matriz identidade. No entanto, é possível demonstrar que, dado que a dimensão d seja grande o suficiente, esses vetores devem ser quase ortogonais [47]. Dessa forma, o produto pode ser decomposto em:

$$RR^T = I + \epsilon$$

Onde $\epsilon_{i,j} = 0$ para $i = j$ e, sendo a matriz R composta por vetores linha $\mathbf{r}$, $\epsilon_{i,j} = \mathbf{r}_i \mathbf{r}_j^T$ quando $i \neq j$. Se a distribuição que gerou os vetores de projeção possui média zero, mostra-se também que $E[\epsilon_{i,j}] = 0$. Além disso, mostra-se que a variância é inversamente proporcional à dimensão do menor espaço [46], ou seja:

$$\sigma_\epsilon^2 \approx \frac{1}{d}$$

A interpretação disso é que a projeção aleatória preserva a similaridade entre os elementos no espaço de busca original, porém introduzido um ruído cuja variância é maior quanto mais agressiva for a redução. Experiências com esse tipo de projeção mostram resultados comparáveis a técnicas como PCA, porém com custo computacional mais baixo [48]. Há também trabalhos utilizando a projeção aleatória como preprocessamento para LSI [49].

3.4 Mapas auto-organizáveis

Os mapas auto-organizáveis (SOM, da sigla em inglês para *Self-Organizing Maps*) são um tipo de rede neural artificial que, tipicamente, utiliza treinamento não supervisionado [50]. O processo de treinamento gera uma representação em baixa dimensão, geralmente em duas dimensões, que mapeia o espaço original, de dimensão elevada. Os neurônios são dispostos sobre uma grade tipicamente retangular ou hexagonal, e uma função de vizinhança é usada durante o treinamento com o intuito de preservar no mapa as características topológicas do espaço de entrada, de dimensão elevada. Na Figura 3.2 pode-se visualizar o diagrama de um mapa auto-organizável bidimensional.

Associado a cada neurônio do mapa há um vetor de pesos da mesma dimensão do espaço de entrada. Durante a fase de treinamento, ao selecionar um elemento do

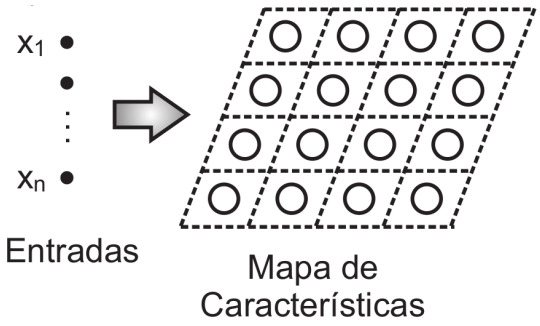


Figura 3.2: Diagrama de um mapa auto-organizável. Retirado de [51]

conjunto de dados, determina-se o neurônio vencedor por aquele cujo vetor de peso está menos distante do vetor de entrada. O treinamento desse neurônio consiste em atualizar seus pesos movendo-o na direção do vetor de entrada. No entanto, os neurônios vizinhos ao vencedor também são treinados, mas com o peso diminuído, de acordo com a função de vizinhança.

A formação do SOM ocorre através de três processos [51]:

Competição Para cada vetor de entrada, há apenas um neurônio vencedor;

Cooperação Neurônios de regiões adjacentes irão ser mover para uma mesma direção;

Adaptação Reforça a resposta do neurônio vencedor, e de seus vizinhos, ao padrão de entrada após o ajuste dos pesos sinápticos.

Uma forma de inspecionar o treinamento do mapa é através da representação *U-Matrix*, que mostra graficamente a distância entre um neurônio e seus vizinhos. Com isso pode-se buscar por agrupamentos nos dados, que devem aparecer no mapa como regiões onde neurônios tem pequena distância entre si, mas grande para os demais. A Figura 3.3 exemplifica essa representação para um mapa treinado com o conjunto de dados clássico de Iris. A faixa escura perto do topo do mapa denota que há uma distância grande entre os neurônios que ela separa. Com isso, espera-se encontrar um agrupamento bem definido no topo do mapa, e outros abaixo dessa faixa.

Desta maneira, o SOM forma um mapa semântico, onde eventos semelhantes são mapeados conjuntamente e os distintos separados. Ao utilizar essa técnica sobre documentos representados por um modelo vetorial, uma pequena distância entre

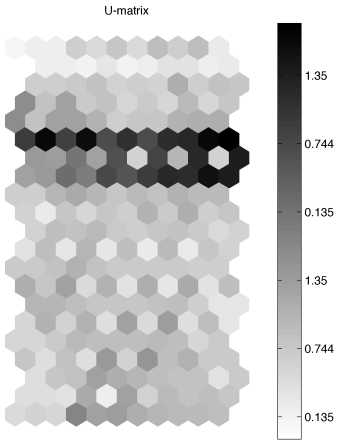


Figura 3.3: Exemplo de representação *U-Matrix*.

dois vetores pode ser interpretada como dois documentos que possuem um conjunto de palavras parecido, o que pode indicar semelhança de assunto.

Mapeamento Topográfico Generativo

O Mapeamento Topográfico Generativo [52] (GTM, do inglês *Generative Topographic Map*) é um método alternativo ao SOM, apresentando princípios estatísticos mais fundamentados que o mapa SOM [51]. Diferentemente do Mapas Auto-Organizáveis, GTM não foi desenvolvida no contexto de redes neurais. Algumas limitações do SOM são superadas, como a falta de uma função de custo e a falta da prova de convergência [53].

GTM consiste em uma mistura de Gaussianas no qual os parâmetros são definidos de forma a maximizar a verossimilhança através do algoritmo *Expectation Maximization* (EM) [54]. Define-se um conjunto de pontos x_i no espaço latente associado a um conjunto de funções base $Y(x)$, no qual mapeia-se o espaço latente continuamente e não-linearmente no espaço de dados. Em conjunto aos parâmetros adaptativos W e β, estabelece-se uma mistura de Gaussianas com centros $Y(Wx_i)$ e uma matriz de covariância dada por $\beta^{-1}I$ [53]. Após inicializar W e β, o treino se dá alternando entre o passo E, onde as probabilidades *a posteriori* são calculadas, e o passo M, onde W e β são reestimadas. A Figura 3.4 apresenta a ideia básica do GTM, o mapeamento do espaço latente no espaço de dados.

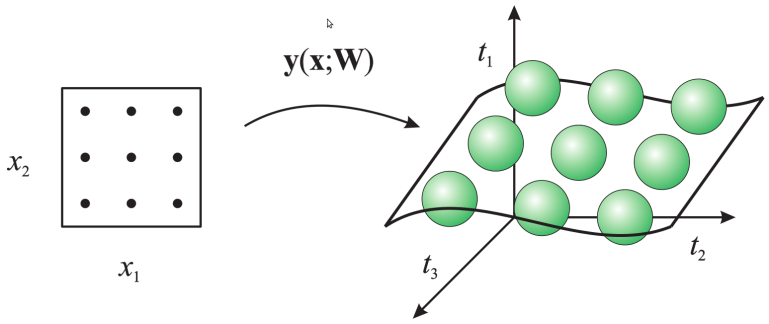


Figura 3.4: Representação do espaço latente mapeado no espaço de dados. Retirado de [52]

3.5 Agrupamento por *K-means*

O algoritmo de *K-means* [55], amplamente utilizado, emprega o conceito de centroides. Dados os K centroides espalhados aleatoriamente no espaço de dados, sendo K igual ao número de agrupamentos pré-definidos, o algoritmo agrupa os eventos, de acordo com o distanciamento entre o evento e o centroide. Forma-se então um diagrama de Voronoi. Como métrica de distância, a mais comum é a distância euclidiana quadrática:

$$d_{ki}^2 = \| x_k - c_i \|^2 \tag{3.1}$$

onde x_k são os eventos do conjunto de dados e c_i são os centroides dos agrupamentos.

A seguir, os centroides são recalculados como o baricentro dos eventos associados aos seus agrupamentos, redefinindo o diagrama de Voronoi. Esse processo é repetido até que os eventos não mudem mudem mais de agrupamento, ou quando um determinado número de iterações do algoritmo for realizado.

Capítulo 4

Metodologia

Este capítulo apresenta o processo de estudo utilizado no presente trabalho.

4.1 Aquisição de dados

As ferramentas usadas pelo CERN para disponibilização de informações, conforme discutido na Seção 2.3.2, são acessíveis via web. No entanto, esses sistemas não oferecem uma opção para recuperar todos os documentos de uma só vez. Para coletar esses dados então é necessário a utilização de um *web crawler*, que é um programa que navega automaticamente pelas páginas do sistema, registrando os documentos para formação do banco de dados e seguindo as ligações hipertextuais para outras páginas do mesmo sistema.

Existem dois requisitos importantes que precisam ser levados em consideração ao utilizar um *crawler* para coletar os dados desses sistemas: ser capaz de navegar por entre as páginas restritas da colaboração e saber extrair o conteúdo em si de cada página, cuja estrutura pode variar de acordo com o sistema sendo acessado.

Como discutido anteriormente, todos os sistemas disponibilizados pelo CERN permitem restringir o acesso ao conteúdo para determinados grupos. Essa restrição pode se tanto para grupos restritos, tais como documentos sigilosos exclusivos a uma área de trabalho, quanto mais abertos como, por exemplo, membros da colaboração. O que observa-se, portanto, é que são poucos os documentos disponibilizados para qualquer pessoa do mundo: a maior parte dos documentos são públicos somente para membros da colaboração. Para ter acesso a esses documentos, é necessário que

o *crawler* saiba lidar com o sistema de autenticação utilizado, o *CERN Single Sign On* [56], e forneça credenciais válidas.

Além disso, ao recuperar uma página é necessário extrair dela o conteúdo bruto, que é o que será usado para as análises, eliminando elementos de navegação e menus. No caso da TWiki, por exemplo, é necessário retirar a barra lateral, os cabeçalhos e rodapés, pois contém informações sobre o sistema da TWiki em si, e não sobre o tópico de interesse. A Figura 4.1 exemplifica as partes que são descartadas e aproveitadas. No caso do CDS, o conteúdo pode ser considerado semi-estruturado, pois para cada documento são exibidos, de forma bem demarcada, seus atributos: título, tipo do documento (como tese, nota interna, etc.), resumo, autores, e outros. No escopo deste estudo, no entanto, o *crawler* desenvolvido foi aplicado somente a páginas TWiki. Todo o conteúdo útil da página é utilizado como um único corpo não estruturado.

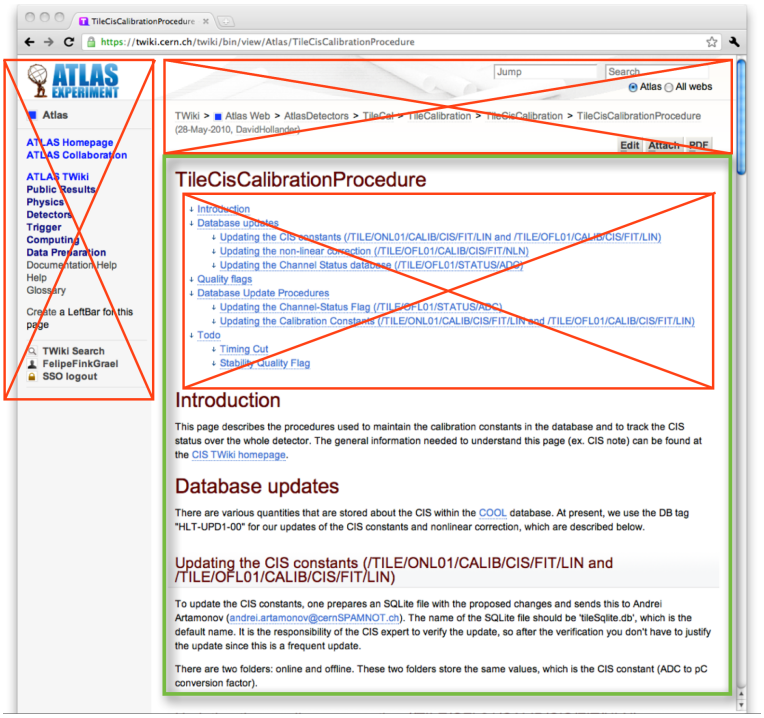


Figura 4.1: Partes de uma página TWiki descartadas na análise.

Existem outros aspectos importantes ao se projetar um *crawler* para uma operação contínua, tais como avaliar o quão frequentemente revisitar uma página a partir da taxa de mudança [57], ou arquiteturas paralelas para recuperação de um grande volume de páginas [58]. No entanto, para o presente trabalho o essencial são os requisitos discutidos acima. Por isso, optou-se por criar um programa *crawler* pró-

prio, que implementa o padrão de autenticação do *CERN Single Sign On* e possui analisadores específicos capazes de extrair o conteúdo útil baseado no documento sendo acessado.

Para recuperar as informações, o *crawler* segue um algoritmo semelhante à busca em largura (*breadth-first search*) [59] em um grafo, onde dado um vértice raiz, visita-se todos os vértices vizinhos antes de visitar outro vértice com profundidade maior. O *crawler* é inicializado com uma lista de endereços URL, que são inseridas em uma lista como nível 1. Ao analisar uma página de um determinado nível, qualquer referência encontrada para outra página do sistema que ainda não tenha sido visitada ou marcada para visitação pelo *crawler* é adicionada ao nível seguinte. Somente após todos os endereços de um nível ser visitados o *crawler* prossegue para o próximo.

4.2 Treinamento dos mapas

Ao analisar as páginas recuperadas pelo *crawler* da seção anterior é possível atribuir manualmente rótulo a algumas delas, pois trazem explicitamente em seu endereço URL a área da colaboração da qual fazem parte. Essa informação será utilizada como informação especialista, para posteriormente avaliar o agrupamento.

Foram gerados dois conjuntos de dados: um contendo todos os dados obtidos e outro contendo somente aqueles que puderam ser rotulados manualmente. Para cada um desses conjuntos de dados foram aplicadas as reduções de dimensionalidade utilizando PCA e RP. Esses conjuntos de dados foram usados para treinar o SOM. Em todos os mapas, a grade usada foi a hexagonal, e a função de vizinhança gaussiana. As dimensões do mapa são determinadas a partir do seguinte critério [60]:

$$\begin{cases} n_x n_y = 5\sqrt{n_d} \\ \frac{n_y}{n_x} = \sqrt{0.75}\frac{\lambda_1}{\lambda_2} \end{cases}$$

Onde n_x é o número de colunas e n_y o de linhas na grade, n_d o número de amostras disponíveis para treinamento e λ_1 e λ_2 os dois maiores autovalores da análise PCA para esse conjunto de dados.

Para agrupar os dados, o algoritmo *k-means* foi executado sobre o *codebook* do mapa, isto é, o conjunto formado pelos vetores de peso dos neurônios do SOM, que

possuem a mesma dimensionalidade dos dados que foram usados para treiná-lo.

4.3 Avaliação do agrupamento

Uma métrica frequentemente usada para avaliar a qualidade de agrupamentos é o índice de Davies-Bouldin [61]. Esse índice avalia ao mesmo tempo tanto o grau de espalhamento dentro de um agrupamento quanto o grau de separação entre eles. O espalhamento de um agrupamento é dado como uma média da distância entre o centroide e os dados correspondentes. O grau de separação entre os agrupamentos é avaliado pelo somatório das distâncias entre os centroides. O índice final é diretamente proporcional ao espalhamento dentro de um agrupamento e inversamente à separação entre agrupamentos, levando em consideração o pior caso. Dessa forma, quanto menor o valor, melhor o agrupamento dos dados.

Para avaliar se os agrupamentos estão ligados às áreas de trabalho do detector, é necessário fazer uso do rotulamento realizado manualmente sobre parte dos dados recuperados. A cada agrupamento é atribuído um rótulo, observando-se a quantidade de documentos rotulados pertencentes a ele. No entanto, devido ao fato dos rótulos não terem o mesmo número de documentos associados, o que é observado não é o número absoluto, mas o percentual sobre o total daquele rótulo. O intuito é facilitar a visualização da região mais importante para um determinado rótulo, mesmo que ele contenha poucos documentos.

Com isso é possível montar uma matriz de confusão, onde cada elemento $c_{i,j}$ corresponde ao número de elementos que pertencem ao rótulo i, mas caíram sobre um agrupamento rotulado como j. A partir dessa matriz são extraídas duas métricas para avaliar o agrupamento: o total de confusão, isto é, o percentual de todos os documentos que foram rotulados errado, e a taxa média de verdadeiros positivos.

Visto que o treinamento é não supervisionado, esses a importância dessas métricas é identificar caso o mapa tenha algum problema para agrupar as páginas por similaridade, em vez de medir uma eficiência de classificação em si.

Capítulo 5

Resultados

O banco de dados utilizado no presente estudo foi montado utilizando o *crawler* descrito na Seção 4.1, aplicado somente às páginas TWiki, tendo a página inicial da colaboração ATLAS como ponto de partida, e seguindo ligações até uma profundidade de quatro níveis, que é suficiente para pegar um conjunto significativo das páginas do experimento, sem adicionar muitas páginas de outras colaborações do CERN. Com isso, foi possível recuperar 16.456 páginas. No entanto, deve-se notar que cada usuário que se cadastra no CERN ganha automaticamente uma página pessoal gerada pela própria TWiki, com instruções de uso do sistema. Os usuários podem alterar esse conteúdo também, mas efetivamente são poucos os que o fazem. Para que essa repetição de um mesmo texto de boas-vindas por diversas páginas não influencie negativamente as análises a seguir, esses documentos foram eliminados antes mesmo da montagem do modelo vetorial. Com isso, restaram para análise 4.643 documentos.

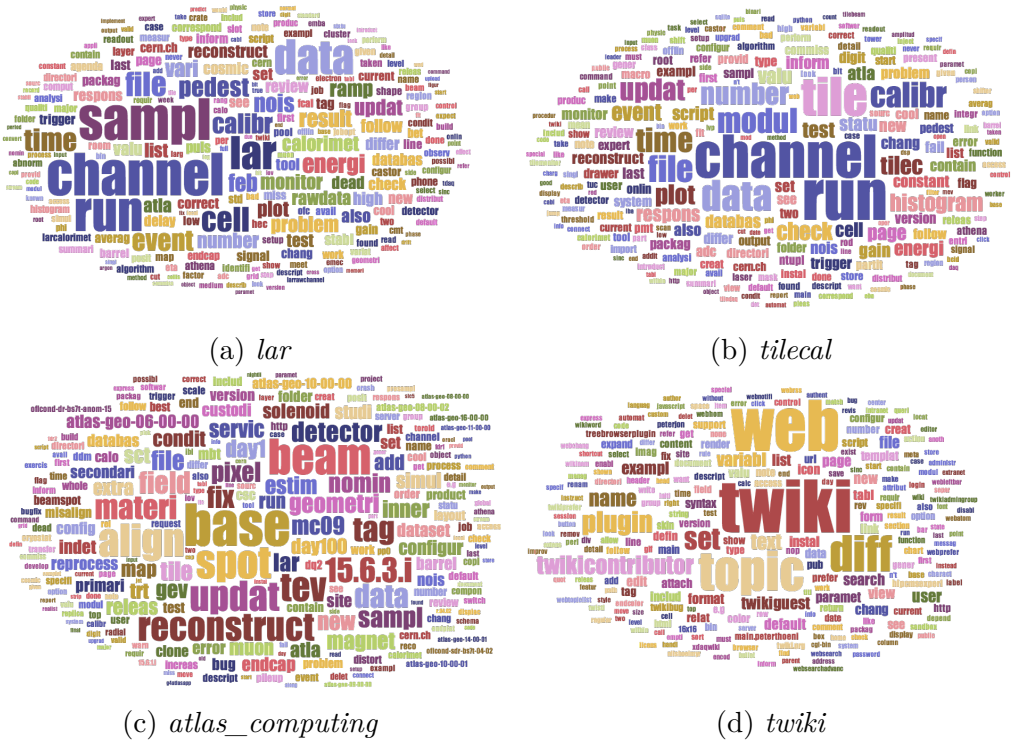
Observando o endereço de cada página recuperada, foi possível rotular um subconjunto dos documentos. A Tabela 5.1 mostra os rótulos resultantes e o número de documentos. Esses rótulos remetem a áreas da colaboração, exceto por *other_cern*, que são páginas fora do experimento ATLAS, e por *twiki*, que são páginas de ajuda do próprio sistema de Wiki, que não fazem parte do CERN. Este último rótulo não foi removido da análise pelo fato de terem conteúdo legítimo, ao contrário das páginas pessoais citadas anteriormente, e para ser usado como controle, pois espera-se que haja pouca confusão entre essas e outras páginas da colaboração.

A Figura 5.4 mostra diagramas de nuvem de palavras formados por documentos

Rótulo	Documentos
atlas_computing	262
tilecal	151
physics_office	15
muon	197
inner_detector	288
tdaq	420
upgrade	81
twiki	194
lar	71
other_cern	53
Total	**1.732**

Tabela 5.1: Rótulos para conjunto de dados recuperados da TWiki

de quatro dos rótulos. Nesse diagrama, o tamanho da palavra é proporcional à sua frequência. Observa-se que os rótulos associados aos calorímetros eletromagnético (*lar*) e hadrônico (*tilecal*) possuem um conjunto de palavras parecido, entre si, mas diferente do rótulo *atlas_computing*. Pode-se ainda observar que o rótulo *twiki* possui, como esperado, um conjunto de palavras bastante diferente dos demais.



(a) *lar* (b) *tilecal*

(c) *atlas_computing* (d) *twiki*

Figura 5.1: Diagramas de nuvem de palavras para documentos rotulados com *lar*, *tilecal*, *atlas_computing* e *twiki*

O conjunto de dos documentos adquiridos foi utilizado por inteiro para definir a representação vetorial das páginas. Uma análise inicial contou um total de 237.831

termos, dos quais 161.483 são *hapax legomena*, ou seja, ocorrem somente uma vez em todo o conjunto de dados. Para a redução de dimensionalidade foi feito um estudo com PCA e projeção aleatória. No entanto, como mesmo após a retirada dos *hapax legomena* a dimensão restante ainda representa um esforço computacional muito alto para a aplicação do PCA, foram escolhidos os 10.000 termos mais frequentes. Isto representa exigir que uma palavra tenha ocorrido em pelo menos oito, ou em cerca de 0,17% dos documentos. Não foi estabelecido um limite máximo para frequência.

Dois conjuntos de dados foram então montados: um contendo somente os documentos rotulados e outro com todos os dados disponíveis. Para redução de dimensionalidade, foi analisada a curva de carga da PCA para esses dois conjuntos de dados. Todos os 10.000 componentes principais foram calculados mas, conforme pode-se observar nas Figuras 5.2 e 5.3, quase toda a energia do sinal fica depositada sobre uma fração pequena do total. Para o conjunto com todos os dados, observa-se que 3.000 componentes são suficientes para absorver 99,3% da energia. No conjunto somente com dados rotulados esse mesmo patamar é atingido com somente 1300 componentes.

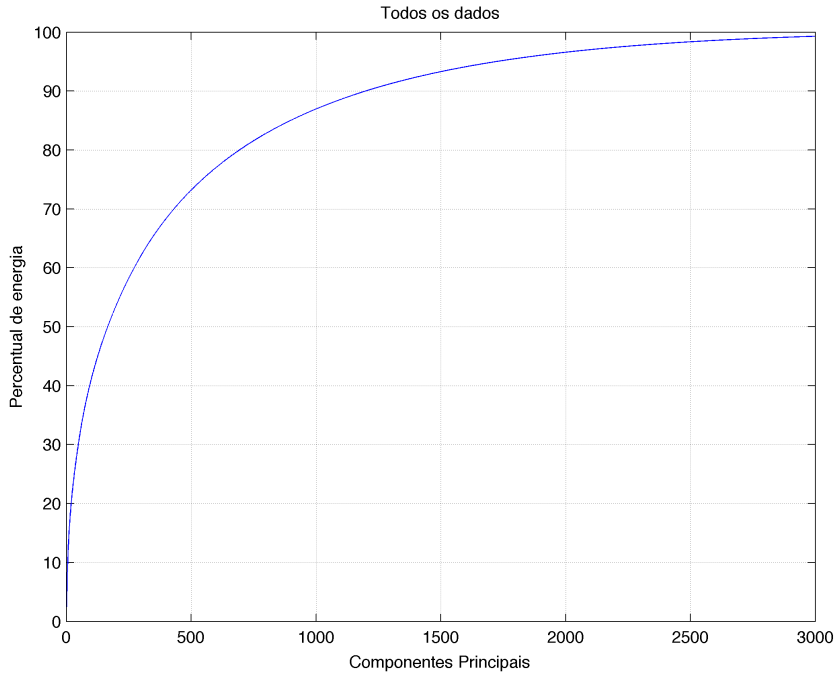


Figura 5.2: Curva de carga da PCA para conjunto com todos os dados

Para a compactação em si foram escolhidos quatro dimensionalidades, que foram escolhidas a fim de representar desde estratégias mais agressivas, admitindo grandes

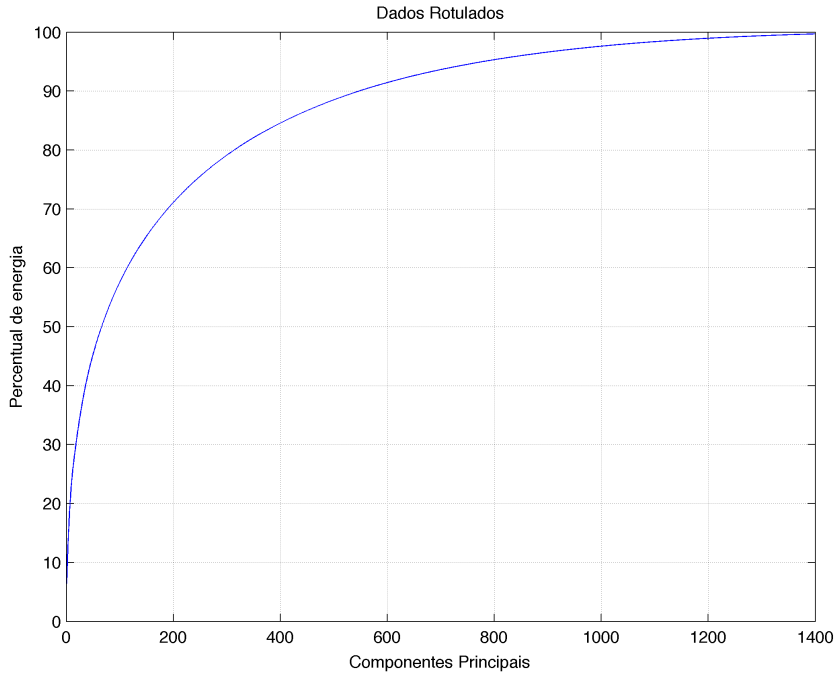


Figura 5.3: Curva de carga da PCA para conjunto somente com dados rotulados

perdas da energia do sinal original, até mais conservadoras: 100, 200, 500 e 1000 componentes. A Tabela 5.2 apresenta a relação entre esses números de componentes e o percentual da energia dos dados originais preservados nessas compressões.

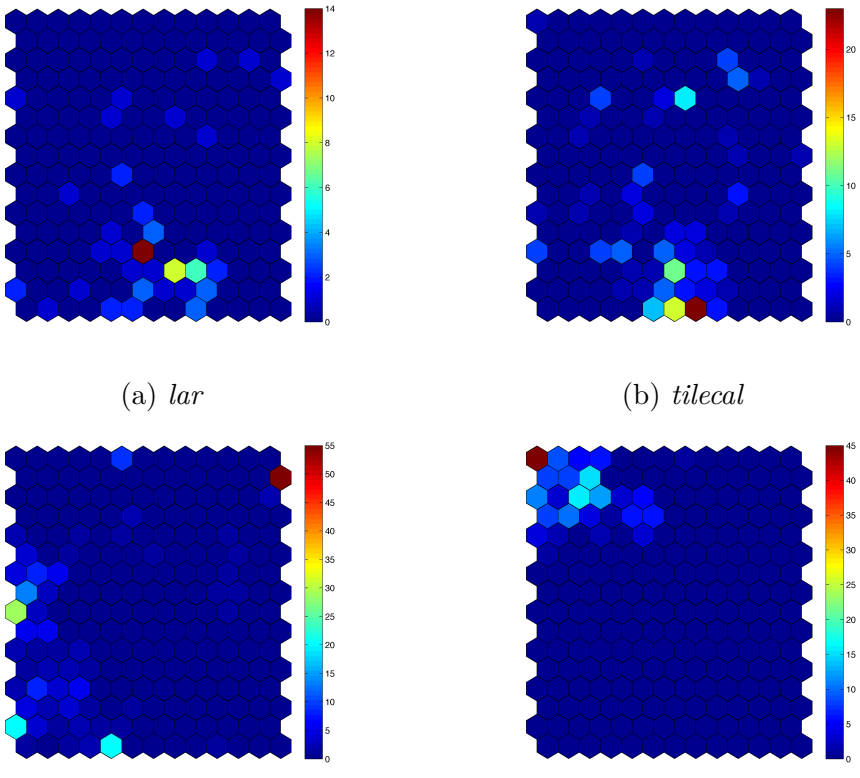
Dimensões	**Todos os Dados**	**Dados Rotulados**
100	41,2%	57,6%
200	53,7%	71,1%
500	73,2%	88,5%
1000	87,0%	97,6%

Tabela 5.2: Percentual da energia do sinal original para número de componentes para compactação com PCA

Para a análise de projeção aleatória, foram gerados mapeamentos onde a dimensão final tivesse o mesmo número de componentes que os usados na análise PCA. Para cada caso foram geradas múltiplas matrizes de projeção, a fim de observar como a aleatoriedade da projeção afeta o resultado final.

Para treinamento dos mapas, os dados foram normalizados para vetores unitários. Os mapas foram treinados utilizando grade hexagonal, com função de vizinhança gaussiana e treinamento por batelada. Os mapas que foram treinado somente com os dados que foram rotulados não tiveram a informação de rotulação agregada à entrada do treinamento, mantendo a característica não supervisionada.

Para exemplificar o resultado dos treinamentos, a Figura 5.4d mostra como documentos dos rótulos *lar*, *tilecal*, *atlas_computing* e *twiki* foram mapeados pelo SOM treinado somente com os dados rotulados, e reduzido para 100 dimensões por PCA. Como esperado, os documentos *twiki* ficaram contidos no canto superior esquerdo do mapa. Os documentos dos calorímetros ficaram em uma região próxima do mapa, se confundindo em alguns casos. Esse comportamento é esperado, visto que as páginas de fato tratam, por vezes, de assuntos em comum dada a similaridade da finalidade desses sub-detectores. Os documentos do rótulo *atlas_computing* Ficaram separados dos demais, se concentrando em um ponto no canto superior direito e em uma região no canto esquerdo, podendo denotar dois subgrupos desse rótulo. Os demais rótulos também ficaram mapeados em regiões concentradas no mapa. O mesmo comportamento se observa para os demais treinamentos com a redução de dimensionalidade por PCA, embora os documentos fiquem ligeiramente menos agrupados à medida que o número de dimensões cresce.



(a) *lar* (b) *tilecal*

(c) *atlas_computing* (d) *twiki*

Figura 5.4: Neurônios ativados por documentos dos rótulos *lar*, *tilecal*, *atlas_computing* e *twiki*

O algoritmo *k-means* foi executado procurando por até 50 agrupamentos sobre os neurônios treinados, e cada configuração foi executada 10 vezes a fim de minimizar o efeito de ótimos locais. O melhor índice Davies-Bouldin obtido foi utilizando a projeção aleatória do espaço original de 10.000 para 500 dimensões, para o conjunto de dados somente com os rotulados. Este resultado sugere que os agrupamentos ficam melhor espalhados nessa configuração. Os demais resultados, no entanto, não estão muito diferentes, sugerindo que, em geral, é possível estabelecer uma separação razoável entre agrupamentos. A Figura 5.5 mostra o resultado desse índice para todos os treinamentos. A barra de erro denota o desvio padrão do resultado para as diferentes matrizes de projeção aleatória criadas com a mesma especificação.

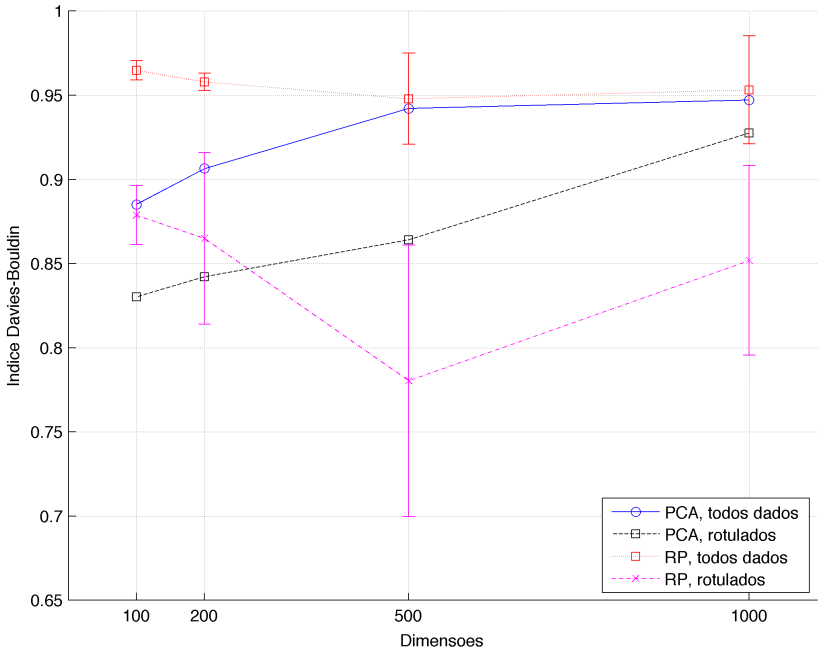


Figura 5.5: Índice Davies-Bouldin para mapas treinados

Ao analisar as taxas de confusão e de verdadeiros positivos, pode-se observar que, como esperado, à medida que o número de dimensões cresce, o resultado por projeção aleatória fica melhor. No entanto, os dados que foram compactados através de PCA mostram desempenho melhor, especialmente nas estratégias mais agressivas de compactação. Observa-se ainda que a diferença de desempenho entre o mapa treinado com todos os dados é pequeno em relação ao treinado somente com os dados rotulados. A Figura 5.6 mostra a taxa de eficiência para os mapas, enquanto a Figura 5.7 mostra a de confusão.

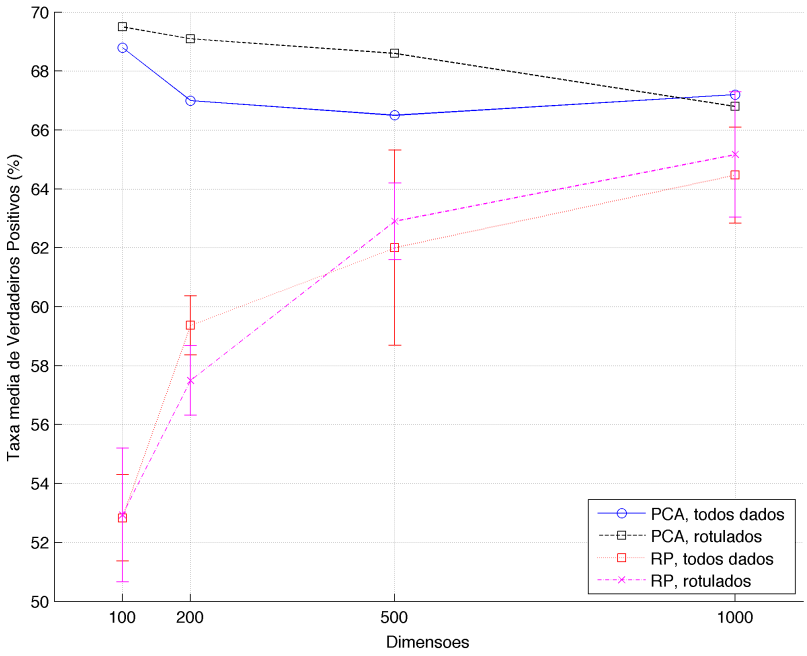


Figura 5.6: Taxa de Verdadeiros Positivos para mapas treinados

Foram treinados também mapas utilizando os dados sem redução de dimensionalidade, ou seja, usando todas as 10.000 dimensões do VSM. Todos os índices de desempenho para esses mapas foram comparáveis, senão piores dos que quando a redução de dimensionalidade foi utilizada. O índice Davies-Bouldin foi o único que apresentou resultados ligeiramente superiores, resultando 0,81 no treinamento com todos os dados e 0,84 para os rotulados. As taxas de verdadeiro positivo ficaram em 64,9% e 67,2%, e a taxa de confusão 40,4% e 36,0% para todos os dados e rotulados, respectivamente.

Tendo esses resultados, podemos focar agora nos mapas treinados usando a compactação por PCA para 100 componentes. No caso do mapa treinado com todos os dados, ao observar como as métricas evoluem de acordo com o número de agrupamentos analisados com o *k-means*, observamos que o melhor valor para todas acontecem para 42 agrupamentos: 68,8% de taxa de verdadeiros positivos, 36,7% de confusão e índice Davies-Bouldin de 0,89. A Figura 5.8 mostra o mapa resultante tanto com os agrupamentos encontrados pelo *k-means* quanto depois de rotulados. Vale notar que essa abordagem permite atribuir o mesmo rótulo a diversos agrupamentos, sendo capaz, portanto, de encontrar agrupamentos maiores de formas irregulares.

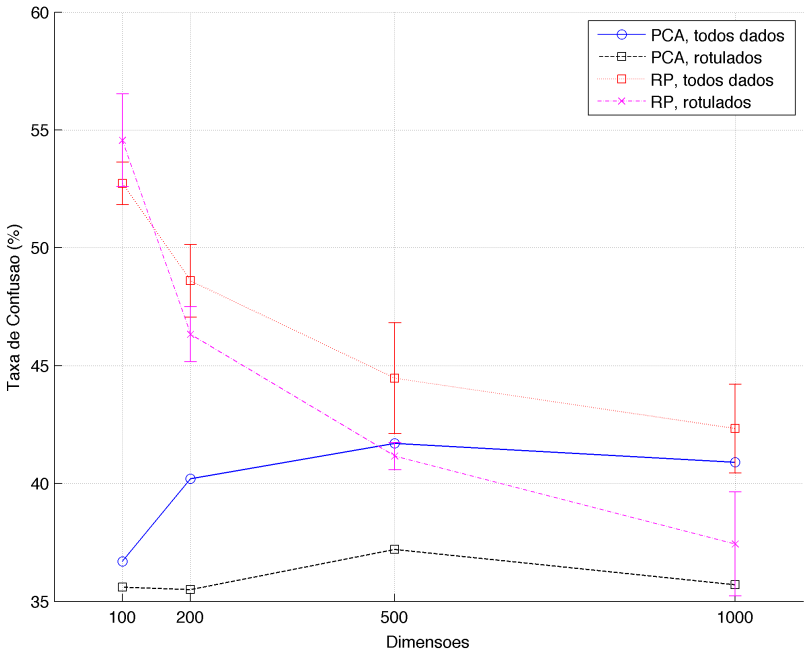


Figura 5.7: Taxa de Confusão para mapas treinados

O mapa treinado somente com os dados rotulados, compactados para 100 componentes por PCA, mostram um comportamento um pouco diferente na variação do número de agrupamentos: o melhor escolha é 43 agrupamentos de acordo com o índice Davies-Bouldin, 42 de acordo com a taxa de confusão, 49 por taxa de verdadeiros positivos. Visto que a taxa de confusão e de verdadeiros positivos não variam muito nessa região, é preferível escolher o de menor índice Davies-Bouldin. Com isso, obtém-se 68% de taxa de verdadeiros positivos, 37% de confusão e um índice Davies-Bouldin de 0,83. A Figura 5.9 Ilustra a evolução desses índices, enquanto a Figura 5.10 mostra como fica o mapa para com 43 agrupamentos.

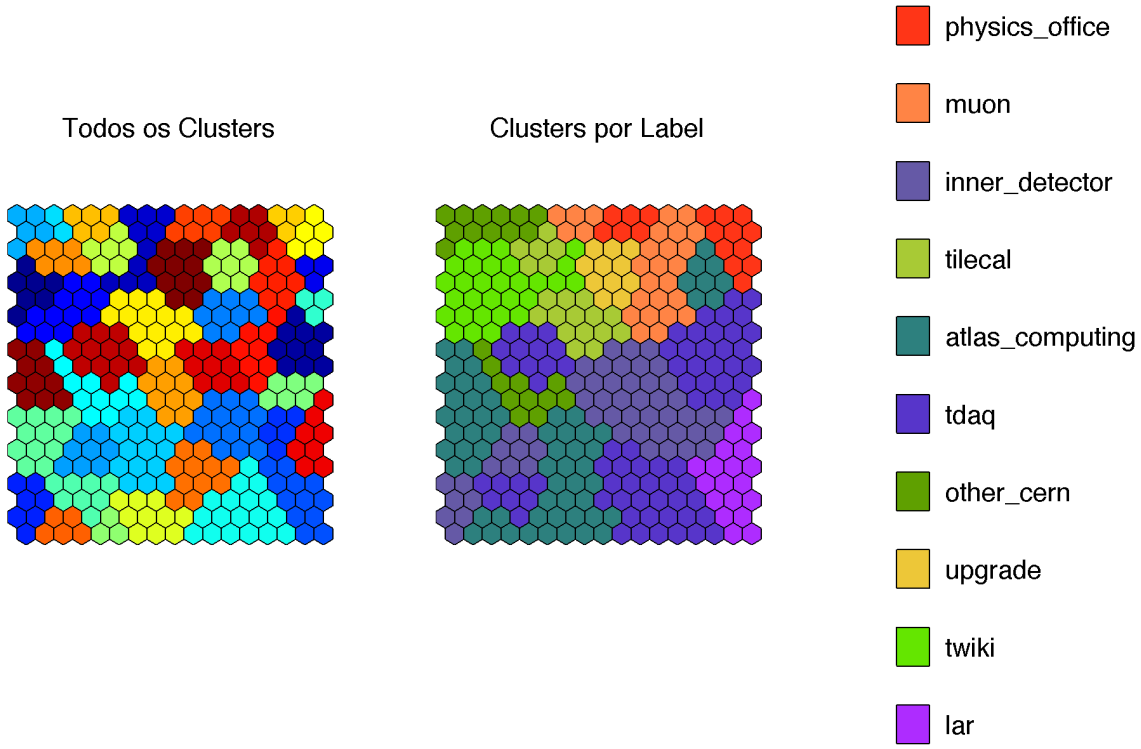


Figura 5.8: Agrupamentos e rótulos para mapa treinado com todos os dados, reduzidos para 100 componentes por PCA

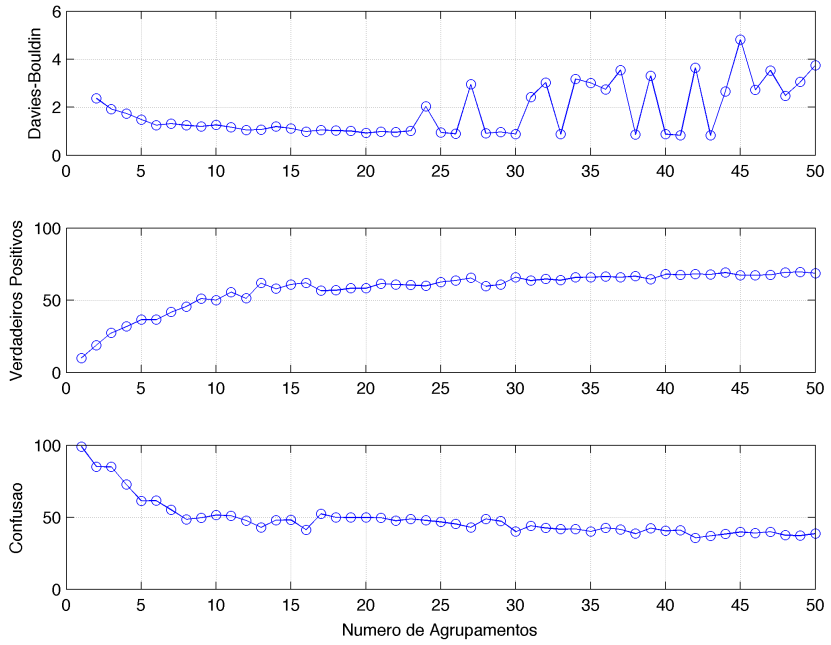


Figura 5.9: Índice de Davies-Bouldin, Taxa de Verdadeiros Positivos e Confusão para dados rotulados, reduzidos para 100 componentes por PCA

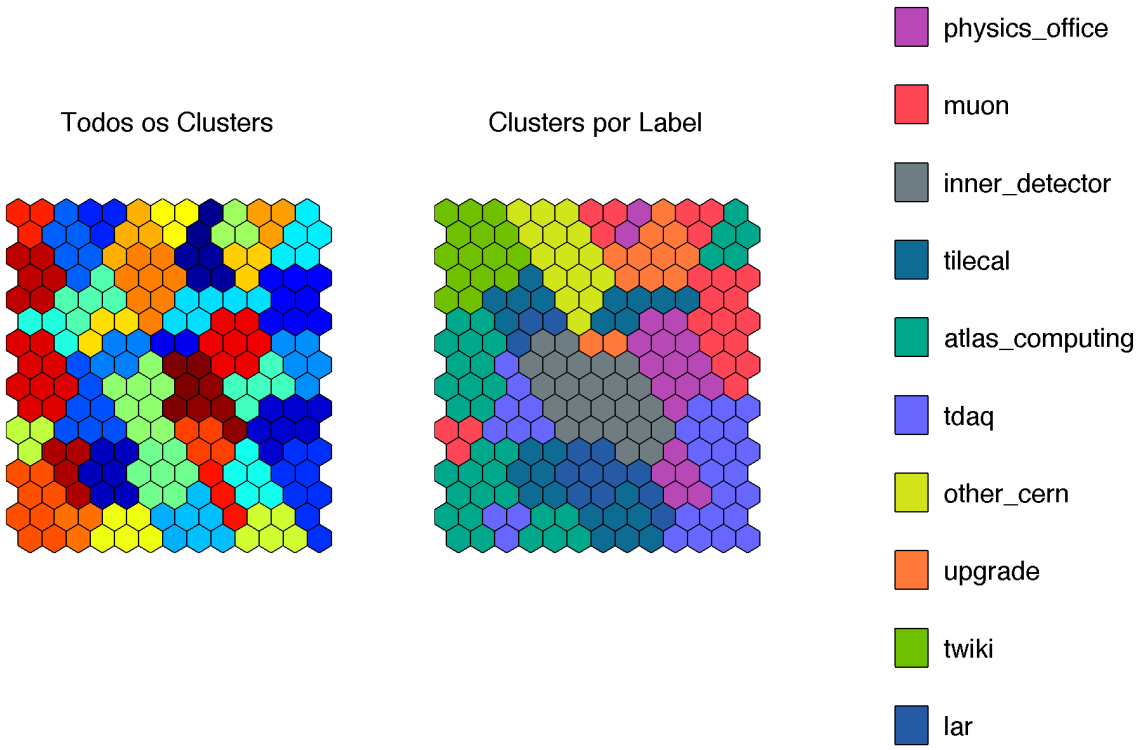


Figura 5.10: Agrupamentos e rótulos para mapa treinado com dados rotulados, reduzidos para 100 componentes por PCA

Capítulo 6

Conclusões

Mais da metade de todas as informações geradas no cotidiano de uma organização, seja ela empresarial, acadêmica ou científica, correspondem a dados não-estruturados. As técnicas de Inteligência Computacional para mineração de dados em texto são, portanto, de grande importância para gerenciamento do conhecimento em organizações que pretendem se manter flexíveis.

O presente trabalho está inserido no contexto da colaboração internacional que construiu e opera o detector de partículas ATLAS, no CERN. Apesar do modelo de gerência possuir poucos níveis hierárquicos, é bastante complexo visto que é composto por 177 institutos de 38 países. A grande rotatividade das pessoas e a grande mobilidade entre áreas tornam a gerência da colaboração complexa, bem como dificulta a entrada de um novo membro. Os conhecimentos acabam fragmentados em diversos documentos, e frequentemente os especialistas em alguns assuntos já estão em outras áreas da colaboração. Não há uma forma fácil de saber quem tem conhecimento sobre que área do detector.

O presente trabalho teve como objetivo recuperar documentos que são inserido nas ferramentas disponibilizadas pelo CERN, como a TWiki, CDS e Indico, e estabelecer um processo através do qual estes sejam agrupados por área da colaboração. Esta é uma etapa importante para gerenciamento da colaboração, visto que permite uma análise posterior das pessoas envolvidas neste documento com sua área de trabalho.

Para conseguir isso, foi desenvolvido nesse trabalho um *crawler* capaz de percorrer as ferramentas web utilizadas no CERN para troca de informações e coletar

dados para análise. Esse *crawler* se comunica com o protocolo de autenticação do CERN, permitindo o acesso a um universo de documentos maior do que simplesmente os completamente públicos. Visto que desta forma o *crawler* tem acesso aos documentos destinado a todos os colaboradores, pode-se ter um melhor panorama da contribuição de cada um.

Esse *crawler* foi aplicado sobre a TWiki, ferramenta na qual páginas Web podem ser escritas de forma colaborativa. Esta é uma ferramenta viva, recebendo mais de 400.000 atualizações por mês, e tendo cerca de mil páginas criadas de Janeiro a Junho de 2013. Com isso foi possível coletar dados para realizar o estudo de agrupamento por áreas da colaboração.

As páginas recuperadas, depois de filtradas para eliminar as que não carregam informações da colaboração, foram representadas vetorialmente usando VSM. As palavras foram lematizadas e, com isso, foram encontrados 237.831 termos, dos quais os 10.000 mais frequentes foram escolhidos para compor o espaço vetorial. Um estudo de redução de dimensionalidade foi conduzido, utilizando PCA e projeção aleatória, com o intuito de evitar os efeitos adversos da alta dimensionalidade e fazer uma seleção de características, descartando componentes que agregam pouco valor aos dados originais, possivelmente atrapalhando as análises.

Uma parte dos dados adquiridos foi rotulada manualmente, aproveitando-se da organização hierárquica visível no endereço URL de algumas das páginas. Foram então gerados dois conjuntos de dados: um contendo todas as páginas e outro contendo só as páginas rotuladas. Esses conjuntos foram então usados para treinar o modelo SOM, usando as diferentes combinações de preprocessamento com PCA e RP estipuladas pelo estudo de redução de dimensionalidade. Uma vez treinado, foi utilizado o algoritmo *k-means* para identificar os agrupamentos do mapa. Foram analisadas três métricas principais: o índice Davies-Bouldin para agrupamento, e a taxa de confusão e verdadeiros positivos que podem ser calculadas ao atribuir rótulos aos grupos.

Como resultado, os mapas treinados com o preprocessamento por PCA foram os que obtiveram os melhores resultados na análise realizada. Os treinados somente com os dados rotulados apresentaram resultados ligeiramente superiores. No entanto, o melhor caso encontrado nas análises realizadas foi o mapa treinado com

todos os dados, compactados para 100 componentes por PCA, apresentando 68,8% de taxa de verdadeiros positivos, e 36,7% de taxa de confusão.

Uma continuação natural deste trabalho é estender a análise para os autores e demais pessoas envolvidas ou citadas nos documentos, de forma a estabelecer a relação entre colaborador e área de trabalho. Isso pode ser aplicado diretamente para a gerência do experimento, permitindo visualizar e avaliar a alocação de pessoas nas diferentes áreas, ou como uma ferramenta que auxilie um novo colaborador a encontrar as informações necessárias e os especialistas nos assuntos de interesse.

As análises sobre os documentos podem ainda ser adaptadas para agregar informações de mais sistemas, extraindo conteúdo, por exemplo, de apresentações e materiais de apoio para reuniões que são inseridos no Indico. Além disso, é interessante também do ponto de vista gerencial, tomar vantagem da alta dinamicidade dessas fontes de dados para realizar uma análise temporal, mostrando como as áreas se comportam ao longo do tempo, ou como se dá o movimento interno de colaboradores.

Referências Bibliográficas

[1] CONSOLI, D. "Text Mining Technology To Support Enterprise Knowledge Management", , n. 30, 2009.

[2] BOISOT, M., NORDBERG, M., YAMI, S., et al. *Collisions and Collaboration:The Organization of Learning in the ATLAS Experiment at the LHC*. Oxford, Reino Unido, OUP Oxford, 2011.

[3] EVANS, L., BRYANT, P. "LHC Machine", *Journal of Instrumentation*, v. 3, n. 08, ago. 2008.

[4] ACZEL, A. D. "Present at the creation: The story of CERN and the Large Hadron Collider", 2010.

[5] "Página *web* do CERN". `http://public.cern.ch`, Novembro 2012.

[6] THE ATLAS COLLABORATION. "The ATLAS Experiment at the CERN Large Hadron Collider", *Journal of Instrumentation*, v. 3, n. 08, Agosto 2008.

[7] BOELAERT, N., BOELAERT, N. "The ATLAS Experiment", pp. 65–81, 2012.

[8] THE CMS COLLABORATION. "The CMS experiment at the CERN LHC", *Journal of Instrumentation*, v. 3, n. 08, Ago 2008.

[9] SZUMLAK, T. "The LHCb experiment", *Acta Physica Polonica. Series B: Elementary Particle Physics, Nuclear Physics, Statistical Physics, Theory of Relativity, Field Theory*, v. 41, n. 7, pp. 1661–1668, 2010.

[10] THE ALICE COLLABORATION. "The ALICE experiment at the CERN LHC", *Journal of Instrumentation*, v. 3, n. 08, ago. 2008.

[11] "ATLAS *Factsheet*". `http://www.atlas.ch/fact-sheets-1-view.html`, Novembro 2012.

[12] "Página *web* do experimento ATLAS". `http://atlas.ch`, Novembro 2012.

[13] FERREIRA, F. *Identificação de Eventos Baseada na Combinação de Detectores de Altas Energias com Diferentes Tecnologias e Segmentações*. Tese de Mestrado, Rio de Janeiro, Brasil, 2012.

[14] THE ATLAS COLLABORATION. "The ATLAS Inner Detector commissioning and calibration", *The European Physical Journal C*, v. 70, n. 3, pp. 787–821, 2010.

[15] ATLAS ELECTROMAGNETIC LIQUID ARGON COLLABORATION. "Construction, assembly and tests of the ATLAS electromagnetic end-cap calorimeters", *Journal of Instrumentation*, v. 3, n. 06, pp. P06002, 2008.

[16] THE ATLAS TILECAL COLLABORATION. "Readiness of the ATLAS Tile Calorimeter for LHC collisions", *The European Physical Journal C-Particles and Fields*, v. 70, n. 4, pp. 1193–1236, 2010.

[17] COLLABORATION, T. A. "ATLAS Muon Spectrometer", 2009.

[18] "Speakers and publication committee organisation and guidelines". 2008. Disponível em: <https://atlas.web.cern.ch/Atlas/private/ATLAS_CB/CB_Approved_Documents/A59_ATLAS_Organisation%5B1.2%5D.pdf>.

[19] GORDON, H. "Publication Issues in Large Scientific Experiments", *NEON Workshop*, 2009.

[20] FRIAS, L. "ATLAS Analysis Papers and Conference Notes". In: *Journal of Physics: Conference Series*, n. 6, p. 062004. IOP Publishing, 2012.

[21] "Indico". Acessado em Junho 2013. Disponível em: <http://indico.cern.ch/>.

[22] LÓPEZ, J. B. G., FERREIRA, J. P., BARON, T. "Indico central – events organisation, ergonomics and collaboration tools integration", *Journal of Physics: Conference Series*, v. 219, n. 8, pp. 082002+, abr. 2010. Disponível em: <http://dx.doi.org/10.1088/1742-6596/219/8/082002>.

[23] "CERN Document Server". Acessado em Junho 2013. Disponível em: <http://cds.cern.ch/>.

[24] PEPE, A., BARON, T., GRACCO, M., et al. "CERN Document Server Software: the integrated digital library", 2005.

[25] "CERN TWiki". Acessado em Junho 2013. Disponível em: <http://cern.ch/twiki>.

[26] AMRAM, N., ANTONELLI, S., HAYWOOD, S., et al. "The use of the TWiki Web in ATLAS", v. 219, n. 8, pp. 082008, 2010.

[27] GRAEL, F. F., MAIDANTCHIK, C., ÉVORA, L. H. R. A., et al. "Glance Information System for ATLAS Management", *Journal of Physics: Conference Series*, v. 331, n. 8, dez. 2011.

[28] "OpenDOAR". Acessado em Junho 2013. Disponível em: <http://www.opendoar.org/>.

[29] "Estatísticas da plataforma CERN TWiki". Acessado em Junho 2013. Disponível em: <https://twiki.cern.ch/twiki/bin/view/Main/CERNTWikiStatistics>.

[30] RUSSOM, P. "BI Search and Text Analytics". 2007. Disponível em: <http://tdwi.org/articles/2007/05/09-what-works/bi-search-and-text-analytics.aspx>.

[31] PAGE, L., BRIN, S., MOTWANI, R., et al. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66, Stanford InfoLab, November 1999. Disponível em: <http://ilpubs.stanford.edu:8090/422/>. Previous number = SIDL-WP-1999-0120.

[32] WERMTER, S., CENTRE, T. I. "Selforganizing classification on the Reuters news corpus". In: *In Proceedings of the 19th International Conference on Computational Linguistics, volume 1. Association of Computing Machinery*, pp. 1086–1092, 2002.

[33] COUSSEMENT, K., POEL, D. V. D. *Integrating the Voice of Customers through Call Center Emails into a Decision Support System for Churn Prediction.* Working papers of faculty of economics and business administration, ghent university, belgium, Ghent University, Faculty of Economics and Business Administration, 2008. Disponível em: <http://econpapers.repec.org/RePEc:rug:rugwps:08/502>.

[34] DE VEL, O., ANDERSON, A., CORNEY, M., et al. "Mining e-mail content for author identification forensics", *SIGMOD Rec.*, v. 30, pp. 55–64, Dezembro 2001. ISSN: 0163-5808. doi: http://doi.acm.org/10.1145/604264.604272.

[35] HONKELA, T., KASKI, S., LAGUS, K., et al. "WEBSOM - Self-Organizing Maps of Document Collections". In: *Neurocomputing*, pp. 101–117, 1997.

[36] KISS, T., STRUNK, J. "Unsupervised Multilingual Sentence Boundary Detection", *Comput. Linguist.*, v. 32, n. 4, pp. 485–525, dez. 2006. ISSN: 0891-2017. doi: 10.1162/coli.2006.32.4.485. Disponível em: <http://dx.doi.org/10.1162/coli.2006.32.4.485>.

[37] MARCUS, M., KIM, G., MARCINKIEWICZ, M. A., et al. "The Penn Treebank: annotating predicate argument structure". In: *Proceedings of the workshop on Human Language Technology*, HLT '94, pp. 114–119, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. ISBN: 1-55860-357-3. doi: 10.3115/1075812.1075835. Disponível em: <http://dx.doi.org/10.3115/1075812.1075835>.

[38] PORTER, M. F. "An algorithm for suffix stripping", *Program*, v. 3, n. 14, pp. 130–137, out. 1980.

[39] BIRD, S. "NLTK: the natural language toolkit". In: *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pp. 69–72, 2006.

[40] SALTON, G., WONG, A., YANG, C. S. "A vector space model for automatic indexing", *Commun. ACM*, v. 18, n. 11, pp. 613–620, nov. 1975.

[41] SALTON, G., BUCKLEY, C. "Term-weighting approaches in automatic text retrieval". In: *INFORMATION PROCESSING AND MANAGEMENT*, pp. 513–523, 1988.

[42] BEZERRA, E., MATTOSO, M., XEXEO, G. "Semi-Supervised Clustering of XML Documents: Getting the Most from Structural Information". In: *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, p. 88, 2006.

[43] HAYKIN, S., NETWORK, N. "A comprehensive foundation", *Neural Networks*, v. 2, 2004.

[44] HYVÄRINEN, A., KARHUNEN, J., OJA, E. *Independent Component Analysis*. Adaptive and Learning Systems for Signal Processing, Communications and Control Series. E.U.A, Wiley, 2004.

[45] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., et al. "Indexing by latent semantic analysis", *Journal Of The American Society For Information Science*, v. 41, n. 6, pp. 391–407, 1990.

[46] KASKI, S. "Dimensionality reduction by random mapping: fast similarity computation for clustering". In: *Neural Networks Proceedings, 1998.*

IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on, v. 1, pp. 413–418 vol.1, 1998. doi: 10.1109/IJCNN.1998.682302.

[47] R., H.-N. "Context vectors: general purpose approximate meaning representations self-organized from raw data", *Computational Intelligence: Imitating Life, IEEE Press*, pp. 43–56, 1994.

[48] BINGHAM, E., MANNILA, H. "Random projection in dimensionality reduction: applications to image and text data". In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pp. 245–250, New York, NY, USA, 2001. ACM. ISBN: 1-58113-391-X. doi: 10.1145/502512.502546.

[49] PAPADIMITRIOU, C. H., TAMAKI, H., RAGHAVAN, P., et al. "Latent semantic indexing: a probabilistic analysis". In: *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, PODS '98, pp. 159–168, New York, NY, USA, 1998. ACM. ISBN: 0-89791-996-3. doi: 10.1145/275487.275505. Disponível em: <`http://doi.acm.org/10.1145/275487.275505`>.

[50] Kohonen, T. (Ed.). *Self-organizing maps*. Secaucus, NJ, USA, Springer-Verlag New York, Inc., 1997. ISBN: 3-540-62017-6.

[51] SIMAS FILHO, E. *Análise Não-Linear de Componentes Independentes para uma Filtragem Online Baseada em Calorimetria de Alta Energia e com Fina Segmentação*. Tese de Doutorado, COPPE/UFRJ, Rio de Janeiro, 2010.

[52] BISHOP, C. M., SVENSÉN, M., WILLIAMS, C. K. "GTM: The generative topographic mapping", *Neural computation*, v. 10, n. 1, pp. 215–234, 1998.

[53] PAMPALK, E. "Limitations of the SOM and the GTM", 2001.

[54] MOON, T. K. "The expectation-maximization algorithm", *Signal Processing Magazine, IEEE*, v. 13, n. 6, pp. 47–60, 1996.

[55] KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., et al. "An efficient k-means clustering algorithm: Analysis and implementation", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 24, n. 7, pp. 881–892, 2002.

[56] ORMANCEY, E. "CERN single sign on solution", *Journal of Physics: Conference Series*, v. 119, n. 8, pp. 082008, 2008.

[57] CHO, J., GARCIA-MOLINA, H., WIDOM, J. “Crawling the Web: Discovery and Maintenance of Large-Scale Web Data”. 2001.

[58] CHO, J., GARCIA-MOLINA, H. “Parallel crawlers”. In: *Proceedings of the 11th international conference on World Wide Web*, WWW ’02, pp. 124–135, New York, NY, USA, 2002. ACM.

[59] KNUTH, D. E. *The art of computer programming, volume 1 (3rd ed.): fundamental algorithms*. Redwood City, CA, USA, Addison Wesley Longman Publishing Co., Inc., 1997.

[60] “SOM implementation in SOM Toolbox”. Julho 2013. Disponível em: <`http://www.cis.hut.fi/somtoolbox/documentation/somalg.shtml`>.

[61] DAVIES, D., BOULDIN, D. “A cluster separation measure”, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, , n. 2, pp. 224–227, 1979.