EMPÍRICABR

EXTRAÇÃO AUTOMÁTICA DE CONHECIMENTO EM DOCUMENTOS TEXTUAIS: UM ESTUDO EXPLORATORIO NO DOMINIO DA SUSTENTABILIDADE

C. M. F. A. RIBEIRO, S. O. QUEIROZ, M. ARAUJO e F. COSTA
Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande de Norte
claudia.ribeiro@ifrn.edu.br

Artigo submetido em novembro/2015 e aceito em dezembro/2015

DOI: 10.15628/empiricabr.2015.3830

RESUMO

Este trabalho dedica-se a analise de ferramentas para a extração automática de conhecimento sobre sustentabilidade, a partir de fontes textuais. Mais especificamente, dedica-se a avaliar a efetividade de tais ferramentas em capturar aspectos relacionados as dimensões ambiental, social e econômica presentes na descrição de projetos sustentáveis. Para tal, foi planejado e executado um experimento com a utilização de três ferramentas de mineração de texto, tendo como base um projeto de agricultura sustentável. Como estratégia complementar de analise, foi utilizada a avaliação humana por meio de técnica de crowdsourcing. Apesar da razoável oferta e maturidade das ferramentas de mineração de texto, os resultados obtidos neste estudo explicitam um limite na capacidade de tais ferramentas em capturar aspectos essenciais ao domínio da sustentabilidade.

PALAVRAS-CHAVE: Extração de Conhecimento, Processamento de Linguagem Natural, Sustentabilidade.

KNOWLEDGE EXTRACTION FROM TEXTUAL SOURCES: AN EXPLORATORY STUDY ON THE SUSTAINABILITY DOMAIN

ABSTRACT

This work is aimed at studying tools for the automatic extraction of knowledge about sustainability, from textual sources. More specifically, it is dedicated to evaluating the effectiveness of such tools in capturing aspects of the so-called three dimensions of sustainability, i.e. environmental, social and economic, from the project description. To this end, an experiment was carried out by using three text mining tools and an agriculture project as data source. As a complementary strategy, it was also used the crowdsourcing technique for human assessment. Despite the reasonable offer and maturity of text mining tools, the results of this study show a limit on the ability of such tools to capture essential aspects of the sustainability domain.

KEYWORDS: Knowledge Extraction, Natural Language Processing, Sustainability

1 INTRODUÇÃO

A extração automática de conhecimento, a partir de documentos textuais na web, tem ganhado a atenção dos pesquisadores, na medida em que tal disponibilidade encontra o limite de assimilação deste conteúdo por seres humanos. Com o aperfeiçoamento das ferramentas de busca, que podem rapidamente referenciar milhares de documentos, o desafio voltou-se para a qualidade da informação obtida.

De forma geral, é possível apontar dois aspectos fundamentais envolvidos na busca de informações na web. O primeiro refere-se ao crescimento horizontal da busca, ou seja, no aumento da quantidade de documentos obtidos. Esta é uma propriedade que está diretamente relacionada ao crescimento da web como ambiente preferencial para a publicação de documentos, e, portanto, deve continuar sendo endereçada pelas ferramentas de busca tradicionais.

O segundo refere-se ao crescimento vertical da busca, que diz respeito à densidade de conhecimento que pode ser encontrado nos documentos. É precisamente sobre esta dimensão que este trabalho trata, cujo objetivo principal é investigar a efetividade dos métodos utilizados por ferramentas atuais de extração e estruturação de conhecimento, notadamente no domínio da sustentabilidade.

Sustentabilidade é uma qualidade que pode ser definida como o equilíbrio entre as dimensões ambiental, social e econômica. Assim, se um dado projeto é dito sustentável, pressupõe-se que sua descrição traz elementos textuais sobre estas três dimensões e possivelmente revele a ênfase dada. Por exemplo, um projeto de energia eólica possivelmente enfatiza aspectos econômicos e um projeto de agricultura sustentável, por sua vez, enfatiza aspectos ambientais.

A despeito da heterogeneidade entre as abordagens usadas por ferramentas atuais, e em particular das três ferramentas estudadas, buscou-se estabelecer uma análise comparativa entre as mesmas, tendo como base um cenário motivacional real. Este cenário explicita o esforço para extrair de forma automática aspectos que caracterizam a sustentabilidade um dado projeto.

Para a realização do experimento de extração automática foi utilizada como entrada o *abstract* de um projeto sobre agricultura sustentável (PORTO, 2008). Tal estratégia apoia-se na premissa de que o *abstract* é um elemento textual que potencialmente encerra a essência de um documento. Assim, procurou-se minimizar o esforço computacional de extração. Adicionalmente, foi utilizada uma avaliação humana complementar, por meio de técnica de *crowdsourcing* (YUEN; KING;LEUNG, 2011). Esta técnica baseia-se no conceito de inteligência coletiva, pela qual uma tarefa complexa pode ser realizada pela colaboração de várias pessoas.

Os benefícios advindos da abordagem utilizada neste trabalho podem servir de base para construção de mecanismos de busca mais inteligentes. Por exemplo, para auxiliar gestores na busca e seleção entre projetos sustentáveis com base no seu foco, seja econômico, ambiental ou social. Ou ainda, na busca de projetos correlatos ou complementares. Por exemplo, projetos sustentáveis de usinas eólicas podem ser combinados com projetos de agricultura orgânica para melhor aproveitamento da área.

Este artigo está estruturado da seguinte forma. Além desta introdução, a seção 2 apresenta os principais fundamentos da extração e representação de conhecimento. A seção 3 apresenta as ferramentas de mineração de texto estudadas e uma analise comparativa dos resultados obtidos

pelo experimento realizado. A seção 4 apresenta a extensão deste estudo com a inclusão de técnica de *crowdsourcing* em complementação as ferramentas de mineração de texto. Por fim, a seção 5 são resumidas as conclusões e discutidos possibilidades de trabalhos futuros.

2 EXTRAÇÃO E REPRESENTAÇÃO DE CONHECIMENTO

O comportamento inteligente é fortemente condicionado ao conhecimento. Esta premissa é válida tanto para decisões humanas quanto para ferramentas computacionais. Assim, a despeito da dificuldade natural em tratar explicitamente o conhecimento humano, há muito se busca meios de fazê-lo de forma automatizada.

Inteligência Artificial (IA) ou Inteligência Computacional é a área da computação que tem se dedicado ao estudo do comportamento inteligente por meios computacionais, sendo a representação de conhecimento e raciocínio automatizado (*Knowledge Representation and Reasoning*) a sub-área de IA que trata como um agente de *software* usa "o que sabe" para "decidir" o que fazer (BRACHMAN e LEVESQUE, 2004).

Muitas abordagens para extração de conhecimento a partir de diferentes fontes tem sido propostas, em especial para a web (KOSALA;BLOCKEEL, 2000). No contexto deste trabalho, existe um interesse particular na extração a partir de fontes textuais, conhecido também por mineração de texto (*Text Mining*). Este processo busca identificar padrões de conhecimento em documentos, diferenciando-se da mineração de dados (*Data Mining*) por utilizar fontes textuais não estruturadas.

Mineração de texto tem sido comumente usado em sistemas para recuperação de informação baseadas em pergunta-resposta, que ajudam o usuário a explorar conteúdos relacionados por tópicos (GUPTA;LEHAL, 2009), busca de informações especificas, análise de dados sobre um grande volume de textos, além de favorecer uma melhor compreensão do conteúdo desses documentos (HAN e KAMBER, 2006). De forma geral, técnicas de mineração de texto possibilitam uma análise de alto nível das características dos textos, permitindo uma análise qualitativa e quantitativa sobre o conteúdo de um ou mais documentos. Tais funcionalidades são essenciais para estruturação de sistemas baseados em conhecimento (KBS, do inglês *Knowledge-based Systems*).

A construção de KBS possui exigência similares a Engenharia de Software em geral, tais como métodos, linguagens e ferramentas especializadas (STUDER; BENJAMINS;FENSEL, 1998). Discutir tais métodos e ferramentas está fora do escopo deste trabalho, contudo, a compreensão dos desafios inerentes ao cenário motivacional deste estudo é necessário apresentar minimamente as características de KBS.

Sistemas baseados em conhecimento comumente são considerados sistemas especialistas, uma vez que simulam modelos cognitivos. Para que tal conhecimento esteja disponível para ser utilizado, por exemplo, em processos de tomada de decisão, não é suficiente capturar o conhecimento, mas é também fundamental estruturá-lo e representá-lo formalmente.

Nos últimos anos tem crescido a utilização de tecnologias da web semântica, notadamente ontologias, para a representação de conhecimento. No domínio da sustentabilidade, ontologias tem sido amplamente utilizadas para modelagem ambiental (VILLA;ATHANASIADIS;RIZZOLI, 2009). Comumente definidas como vocabulários compartilhados, as ontologias permitem a formalização de conceitos e de relações entre estes conceitos, além de regras que auxiliam o raciocínio automatizado (WIMALASURIYA;DOU, 2010).

O processo de extração e representação semântica de conhecimento em geral incorre sobre os conceitos (substantivos) identificados no texto. Conceitos representam eventos e objetos do mundo real, que ajudam o usuário a explorar, examinar e entender ideias, ideologias, tendências, pensamentos, opiniões, textos, documentos, livros, mensagens, etc. (LOH *et al*. 2001). Este processo não trivial consiste em cinco fases, conforme ilustrado na Figura 1:

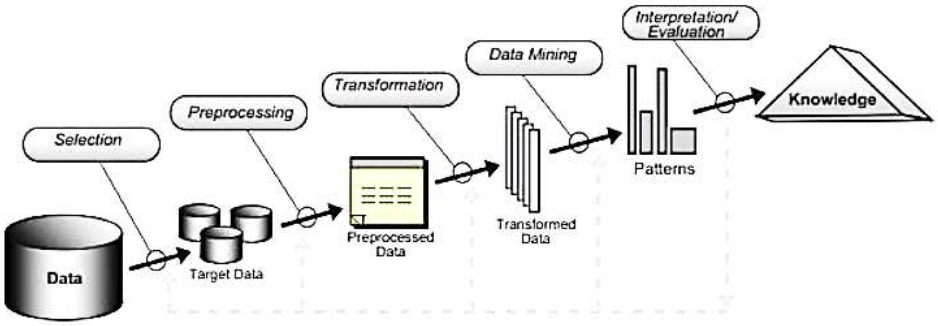

Figura 1: Visão geral das etapas de um processo de extração
Fonte : FAYYAD *et al*., 1996.

- Seleção (*Selection*): dedicada a compreensão do domínio da aplicação, esta etapa é caracterizada por incluir o conhecimento prévio relevante e os objetivos do usuário na extração do conhecimento.
- Pre-Processamento (*Preprocessing*): dedicada a criação de um conjunto de dados para analisar, esta etapa é caracterizada por selecionar um conjunto de dados que serão usados para realizar tarefas de descoberta.
- Transformação (*Transformation*): etapa onde operações básicas de limpeza sobre os dados são realizadas, tais como remoção de ruído e verificação de inconsistência nos dados.
- Mineração de dados (*Data mining*): etapa onde os dados minerados são combinados e confrontados com os objetivos definidos no domínio da aplicação.
- Interpretação (*Evaluation*): Etapa final em que consiste de incorporar os padrões que foram descobertos.

A construção de uma base de conhecimento é geralmente o objetivo final do processo de extração de conhecimento, conforme descrito. Muitas ferramentas foram desenvolvidas para a automatização deste processo.

Para este estudo foram investigadas três ferramentas, selecionadas por sua disponibilidade para uso, documentação e tratamento semântico: GATE, FRED e data-TEXT. O detalhamento das características essenciais destas ferramentas é objeto da seção 3.

3 ANALISE DE FERRAMENTAS DE EXTRAÇÃO DE CONHECIMENTO

Este estudo exploratório foi baseado na mineração textual de um projeto de agricultura sustentável (PORTO, 2008). A sustentabilidade ambiental deste projeto de cultivo consorciado de cenoura e rúcula deve-se à utilização de pesticidas de baixo impacto, bem como a prevenção de esgotamento do solo pela combinação de culturas com raízes de tamanho distintos. Na dimensão econômica este projeto prevê múltiplas colheitas no ano e no aspecto social, a pratica de agricultura familiar, que apresenta dentre outros benefícios a fixação do homem no campo.

Tais características deste projeto sustentável, que foi objeto de tese de doutoramento, estão de certo modo explicitados em sua descrição. Neste estudo, a mineração textual tem como base o *abstract* do documento, por ser um elemento textual auto-contido e conciso, embora completo o suficiente para que a essência do projeto seja capturada. Para fins de simplificação, este ramento textual será referenciado como “texto-base”.

O experimento realizado teve como objetivo principal avaliar a capacidade de extração de conhecimento das ferramentas de mineração utilizadas. Conforme discutido na seção 2, estas ferramentas buscam palavras por meio de análise morfológica, especialmente por conceitos (substantivos) e comumente em língua inglesa. Apesar da significativa variação de técnicas e abordagens, buscou-se estabelecer como parâmetro de analise das ferramentas, a quantidade e relevância de termos relacionados a sustentabilidade obtidas pelas mesmas.

3.1 GATE

GATE é um *framework* Java aberto projetado para o desenvolvimento de componentes para o processamento de linguagem natural. Ao longo de mais de uma década, muitos componentes foram sendo incorporados ao GATE, que é considerada uma das ferramentas mais utilizadas para mineração de texto.

O processo de extração no GATE é caracterizado pelas anotações feitas por uma gramática definida a um documento, onde os conceitos são representados pelos substantivos e usados para construir ontologias. A ênfase desta ferramenta está em anotações semânticas de *corpus*, permitindo que essas anotações sejam feitas de forma manual e/ou automática, podendo ser corrigido os resultados no final do processamento.

Foram realizados diversos testes com o GATE. A eficiência desta ferramenta é comprovada, notadamente na identificação de elementos tais como datas, localidades e outros, onde tais facilidades estão estruturadas em plug-ins que segmentam gradativamente as frases do texto, buscando separar as palavras e identificando sua classe gramatical.

Para efeito deste experimento, contudo, estas funcionalidades não são suficientes, uma vez que buscou-se avaliar a capacidade de extrair conceitos relacionados ao domínio da sustentabilidade. Por exemplo, deseja-se que conceito tais como “pesticida” e “adubo”, sejam automaticamente reconhecidos como sendo conceitos relacionados a esfera ambiental. Para tal, é necessária a utilização de fontes externas, tais como dicionários, que permitam calcular a distância ou similaridade semântica entre conceitos. Durante este estudo, não foi identificado nenhum plug-in que oferecesse esta funcionalidade. A geração automática de ontologias para a representação do conhecimento extraído também mostrou-se insatisfatória.

A Figura 2 mostra a interface e um dos resultados de extração gerado pelo GATE, tendo como entrada o texto-base.

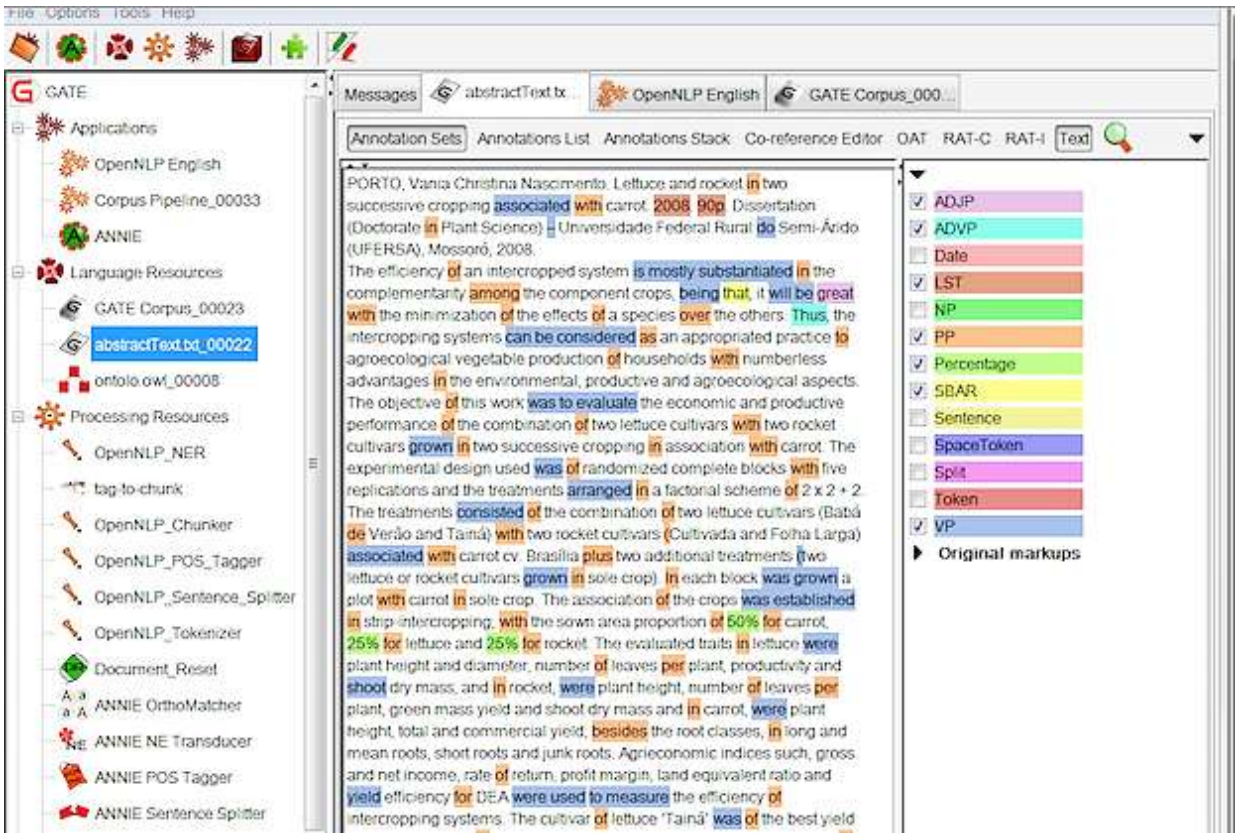

Figura 2: Mineração de texto com GATE.

3.2 FRED

FRED é uma ferramenta que produz automaticamente ontologias de conhecimento extraído de textos e dados vinculados de sentenças em linguagem natural. Baseado em análise semântica, o processo de extração realizada pelo FRED possibilita capturar termos baseados em eventos, através de uma associação dos conceitos extraídos do texto com conceitos de entidades externa.

Os conceitos extraídos pelo FRED são mostrados em hierarquias juntando substantivo e adjetivo, como por exemplo, o conceito "*intercroppedsystem*" que são subclasses "*System*" e "*Intercropped*". Esta ferramenta acrescenta automaticamente na ontologia gerada, diversos outros conceitos, o que o torna propenso a "explosão" de conceitos.

Esta característica praticamente inviabilizou sua utilização com o texto-base. Sendo sua funcionalidade normalmente demonstrada através de frases, durante a realização do experimento foi utilizada somente uma frase do texto-base: "*The efficiency of an intercropped system is mostly substantiated in the complementarity among the component crops*".

A Figura 3 representa a saída gráfica resultante, onde é possível visualizar todas as interconexões entre conceitos e observar que com uma sentença composta de 16 palavras, o grafo gerado contém muitos conceitos, tornando difícil o entendimento.

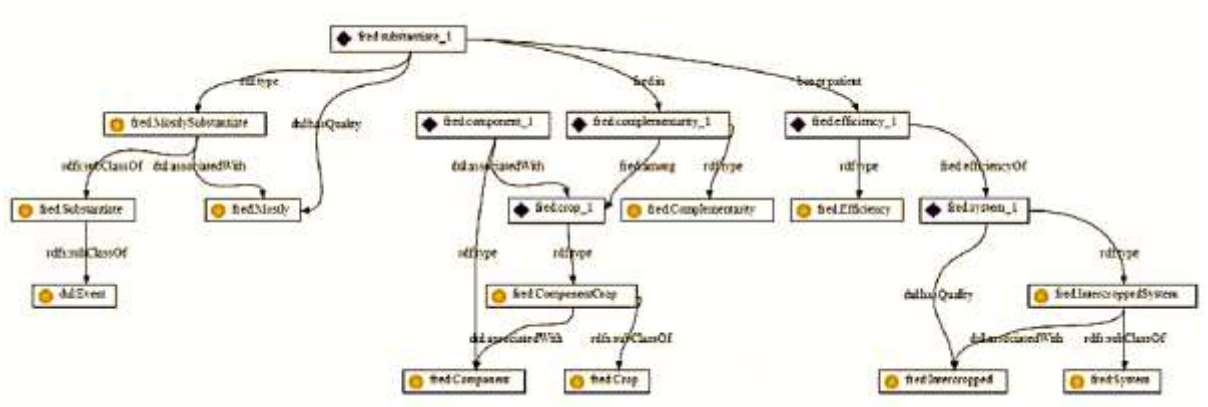

Figura 3: Mineração de texto com FRED

3.3 dataTXT

DataTXT é uma ferramenta composta por um conjunto de APIs de análise semântica de texto, integrada em sistemas existentes através da API REST. O conjunto de APIs inclui o data TXT-NEX (*Named Entity eXtraction*) que realiza extração de conceitos de textos e os liga com dicionários externos e TXT-CL (*Classifier*) que classifica sentenças.

Esta ferramenta pode ser usada com entrada direta de texto ou pela indicação de URL, como pode ser visualizado pela Figura 4, o que possibilita sua utilização para extração de conhecimento de páginas na web. A interface também possui um botão para ajuste da precisão e detecção automática de linguagem, tendo sido reportado a constante atualização de suporte a outros idiomas além do inglês.

Durante o experimento foi usada a API dataTXT-NEX com o texto-base. Esta ferramenta mostrou-se a mais adaptada para o tipo de extração que se deseja obter, o que pode ser observado através do conjunto de conceitos, denominado entidades (*entities*).

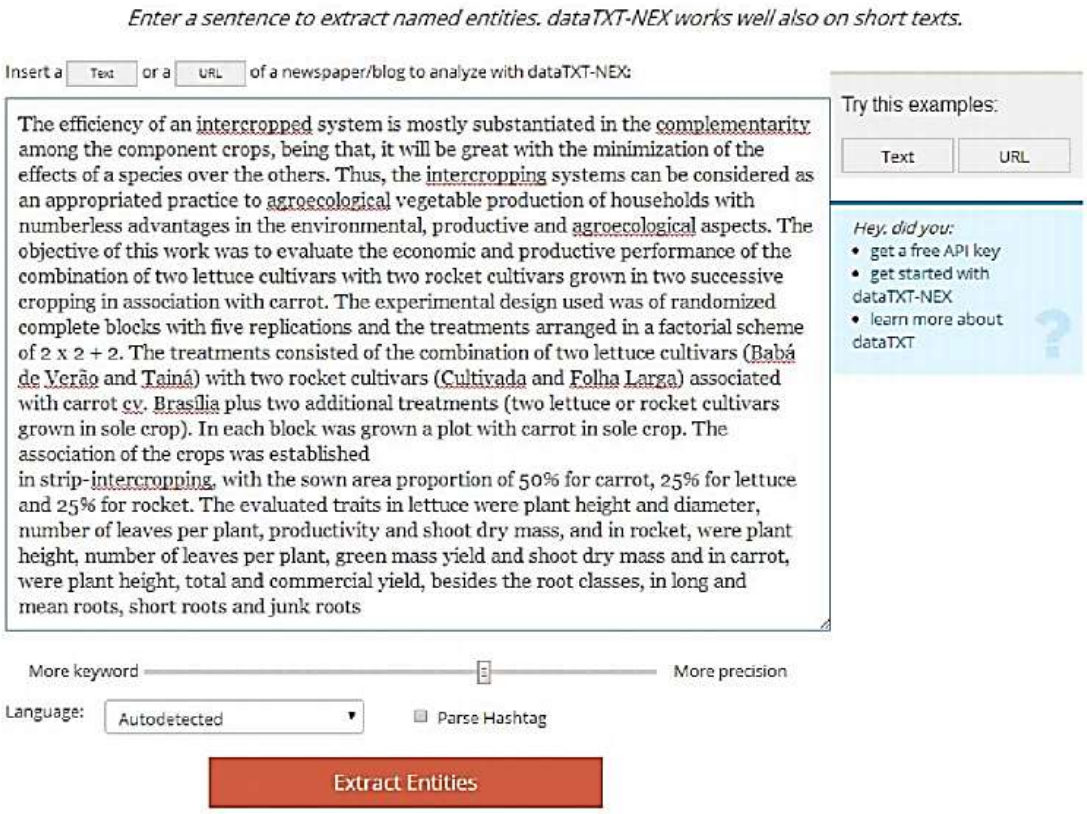

Figura 4: Interface do dataTXT-NEX

Durante o experimento realizado, o dataTXT-NEX extraiu 22 entidades, sendo 2 referentes a lugares e 19 conceitos. Com base na Figura 5, que representa a interface da ferramenta, é possível visualizar através da barra inferior, a saída da mineração do texto-base.

Regionalismos tais como “Babá de Verão” e “Tainá”, nomes locais de duas cultivares referenciadas no experimento, não foram compreendidos pela ferramenta, o que era esperado, estando as mesmas, portanto, fora do conjunto de 19 conceitos identificados.

A saída do dataTXT-NEX é estruturada no formato JSON, sendo os conceitos agrupados de forma automática em categorias, quais sejam “*persons*”, “*work*”, “*organisations*”, “*places*”, “*events*” e “*concepts*”, visualizadas na parte inferior da ferramenta, conforme Figura 5. Esta categorização pretende facilitar o entendimento do trabalho executado pela ferramenta.

3.4 Análise Comparativa

As três ferramentas analisadas neste estudo, quais sejam GATE, FRED e dataTXT, guardam entre si similaridades. A documentação destas ferramentas nem sempre disponibiliza de forma clara, detalhes técnicos sobre os algoritmos de extração e análise de elementos textuais. Contudo, esta ausência não compromete este estudo, que visa objetivamente analisar a capacidade de extração automática de termos que melhor expressem a sustentabilidade do projeto.

Neste aspecto, nenhuma das ferramentas estudadas atendeu de forma significativa este requisito de extração contextualizada, que é fundamental para a construção de mecanismos de busca mais inteligentes. Provavelmente esta lacuna deve-se de alguma forma, a limitação das

ferramentas em usar de forma efetiva as fontes externas, tais como dicionários e ontologias de domínio.

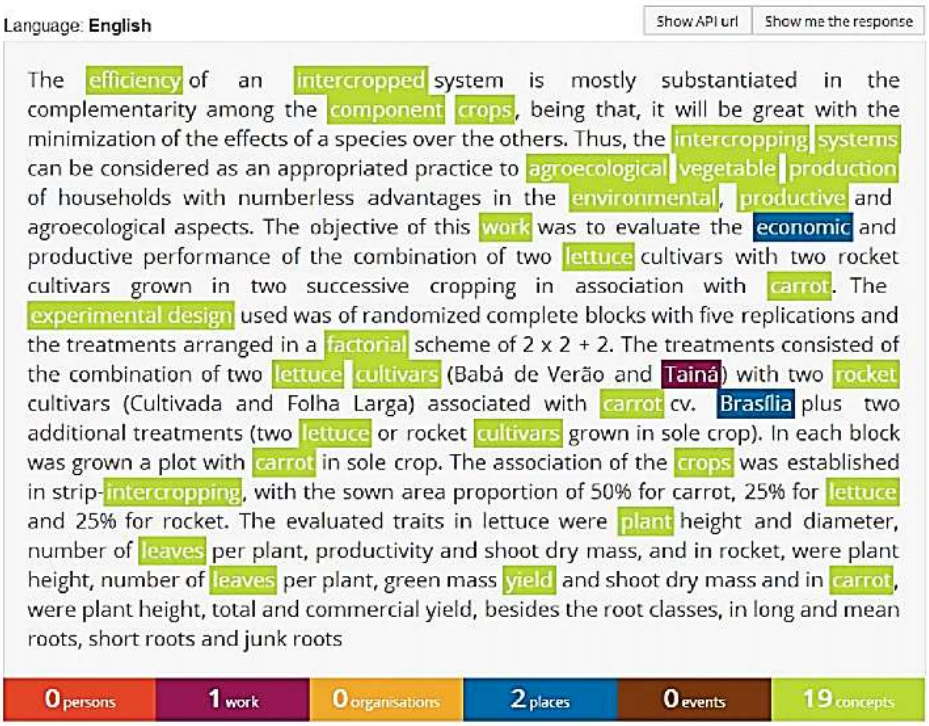

Figura 5: Mineração de texto com data-TXT

Dada a identificação desta limitação de extração contextualizada ou qualificada das ferramentas estudadas, este estudo avançou na busca de técnicas complementares para validação da relevância de termos extraídos por estas ferramentas. Diversas técnicas de avaliação de ferramentas de mineração de texto tem sido propostas, dentre elas as que exploram técnicas de clusterização de termos utilizadas por estas ferramentas (SRIVASTAVA;SAHAMI, 2010). Dentro deste estudo, foi realizada também uma busca por ferramentas adicionais de clusterizacão, que permitissem obter o agrupamento dos termos por esfera da sustentabilidade: ambiental, social e econômica. Não foi identificada nenhuma ferramenta que atendesse razoavelmente a este propósito.

Neste trabalho, optou-se por uma abordagem baseada na percepção humana. De fato, parece existir um consenso sobre o valor que a avaliação de um especialista ainda exerce na analise de qualidade dos processos automatizados. É razoável supor que tal fato aponta para os limites do tratamento automatizado da complexidade inerente a modelagem do conhecimento.

Este trabalho assume a necessidade de complementaridade entre técnicas de mineração de texto automatizadas, tal como fornecida pelas ferramentas estudadas, e a colaboração humana, como forma de ampliar a qualidade dos resultados de extração contextualizada. Para tal, investigou-se a utilização de técnica de *crowdsourcing* (SARASUA; SIMPERL; NOY, 2010). Detalhes da utilização desta abordagem no contexto deste estudo são detalhados a seguir.

4 MINERAÇAO DE TEXTO E CROWDSOURCING

O termo *crowdsourcing* foi utilizado por Jeff Howe e Mark Robinson em 2006, para descrever um novo modelo de negócio baseado na web, que se aproveita da colaboração individual em redes largamente distribuídas. Também chamado de trabalho colaborativo, *crowdsourcing* prevê formas distintas de estimulo a colaboração, remunerada ou não, para a conclusão de uma determinada tarefa.

Diversas categorias de aplicações de *crowdsourcing* tem sido descrita na literatura, sendo as quatro principais: sistemas de votação, sistemas de compartilhamento de informações, jogos e sistemas criativos (YUEN; KING; LEUNG, 2011). Sistemas de votação apresentam diversas tarefas que o colaborador pode escolher, a exemplo do site MTurk (*Amazon Mechanical Turk*). Neste grande mercado virtual, opções de trabalho colaborativo são ofertados de diversas formas.

Em sistemas de votação a resposta correta em geral é obtida por critério de maioria. Assim, é fundamental obter o máximo de adesão, sendo muitas vezes usados além de remuneração, critérios de adesão a causas sociais. São exemplos de sistemas de votação senso comum, sugestões, identificação e nomeação de entidades, avaliação de imagens, avaliação de relevância, anotação de linguagem natural, dentre outros.

No contexto deste estudo, *crowdsourcing* foi utilizado para avaliação de relevância, com o objetivo de validar o conjunto de termos extraídos automaticamente pelas ferramentas estudadas. Como cada ferramenta gerou conjuntos distintos de termos extraídos, procurou-se obter a percepção das pessoas com relação aos termos que melhor descrevem o projeto, a partir do mesmo texto-base usado no experimento.

Apesar desta abordagem requerer maior esforço do participante, que necessita ler o texto e incluir manualmente os termos, procurou-se evitar a potencial indução que um sistema de votação pela apresentação de palavras pode apresentar. Foi desenvolvido então um formulário web para esta finalidade, que inicia com a caracterização sobre grau de escolaridade do participantes. O objetivo é subsidiar posteriores analises sobre a influência do grau de escolaridade sobre a percepção do usuários sobre a temática sustentabilidade.

Ao todo 51 pessoas aceitaram participar do experimento, cujo grau de escolaridade pode ser visualizado pela Figura 6. Como pode ser observado, houve predominância da participação de alunos com graduação em andamento (43%), seguido de 31% por pessoas com ensino médio completo. Alunos de pós-graduação também participaram, embora em menor quantidade.

Nível de Escolaridade	Participantes	%
Ensino Médio Completo	16	31
Graduação em andamento	22	43
Graduação	5	10
Pós-Graduação em andamento	2	4
Pós-Graduação	6	12

Figura 6: Numero de participantes por grau de escolaridade.

Além do texto-base, o formulário também inclui uma explicação breve sobre o propósito do experimento, conforme Figura 7. Além das funcionalidades relacionadas a inclusão/exclusão de

termos considerados mais relevantes pelo participante, foi disponibilizada uma forma básica de clusterização, por meio da indicação da categoria que o participante julga mais adequada para o termo inserido.

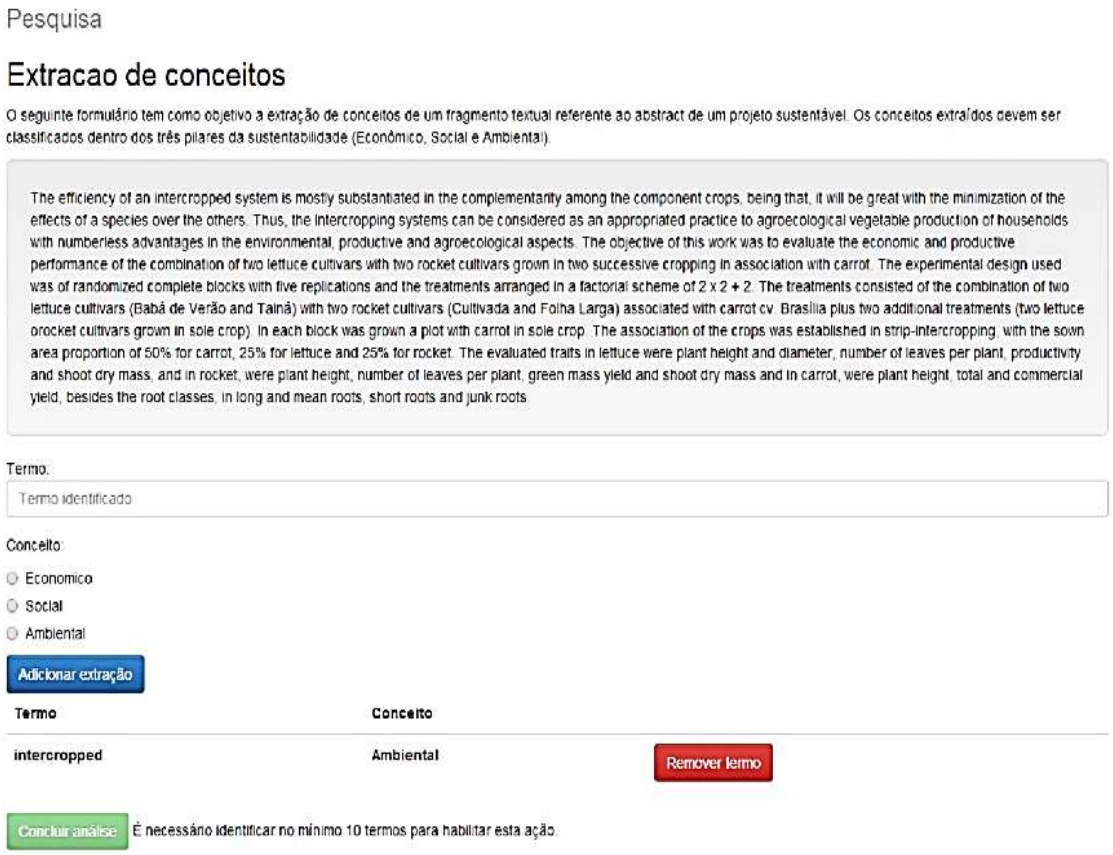

Figura 7: Formulário web para crowdsourcing de analise de relevância

Com base nos dados coletados, foram feitas diversas analises. Por exemplo, identificou-se uma participação mínima de 10 termos por participantes, independente do grau acadêmico do participante. Outra análise efetuada relaciona-se a distribuição dos termos dentro das dimensões da sustentabilidade, considerando a área de estudo dos participantes. Por esta analise, demonstrada pela Figura 8, é possível observar tendência a inclusão de termos relacionados a esfera ambiental, entre os participantes das ciências exatas (56%), engenharias (49%) e ciências biológicas (50%).

Analisando a distribuição de termos por dimensões da sustentabilidade e por nível de escolaridade, não observou-se tendências. Em geral, participantes de todos os níveis incluíram em sua maioria termos categorizados na dimensão “Ambiental”, seguido de termos relacionados ao domínio “Econômico” e por último da dimensão “Social”. Estes dados apontam para a independência com relação ao nível de escolaridade dos participantes.

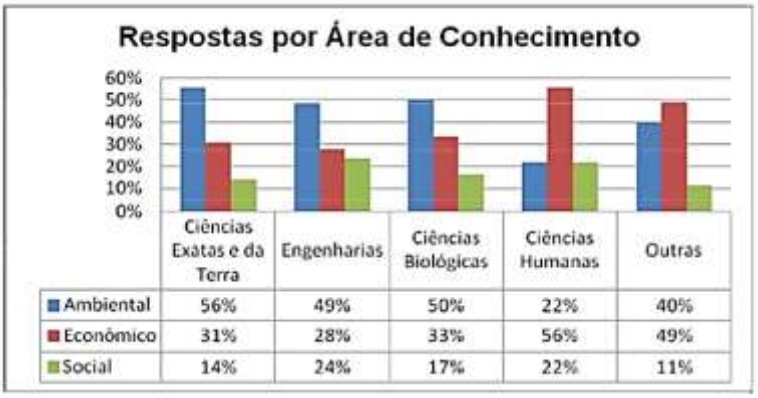

Figura 8: Distribuição dos termos por dimensões de sustentabilidade e área de estudo

Fazendo uma análise comparativa dos resultados da extração automática pela ferramenta dataTXT, considerada a que produziu resultados mais significativos dentre as três ferramentas analisadas, e os resultados obtidos pelos participantes na aplicação *crowdsourcing*, observou-se uma interseção de 90% dos termos, ou seja, de verdadeiros positivos. Tal resultado pode ser interpretado como alta efetividade da ferramenta em capturar os termos de maior relevância para a caracterização do projeto sustentável. Os termos que caracterizam regionalismos, não foram considerados como relevantes nem pela ferramenta e nem pelos participantes, o que reforça a efetividade da ferramenta dataTXT.

Do experimento realizado, vale destacar a explicitação da importância e adequabilidade da técnica de *crowdsourcing* como estratégia complementar para validação de ferramentas de mineração de texto. Tal constatação aponta na direção de novos trabalhos, que são discutidos a seguir juntamente com as principais conclusões deste trabalho.

5 CONCLUSÕES E TRABALHOS FUTUROS

O auxilio computacional para analise de textos na web torna-se cada vez mais necessário. Não somente pela quantidade de documentos disponibilizados, mas principalmente pela natureza e densidade dos documentos.

Neste trabalho, foram investigadas três importantes ferramentas disponíveis para mineração de texto: GATE, FRED e dataTXT. A escolha destas ferramentas foi orientada por critérios de utilização de tecnologias semânticas, notadamente ontologias, para a estruturação do conhecimento obtido por estas ferramentas. Ontologias são atualmente consideradas um valioso instrumento para representação de conhecimento e raciocínio automatizado.

O experimento realizado teve como base um fragmento textual de uma tese de doutorado sobre projeto de agricultura sustentável. O fragmento, no caso, foi o *abstract*, pelo mesmo sintetizar as principais características do estudo. A estratégia permitiu avaliar a efetividade das ferramentas de mineração de texto de uma perspectiva pragmática, em um cenário real de busca por projetos sustentáveis.

Das ferramentas analisadas, o dataTXT apresentou o conjunto de termos extraídos considerados de maior relevância para o domínio da sustentabilidade, conforme objetivo principal deste estudo. A aferição desta efetividade foi ampliada pela utilização de técnica de *crowdsourcing*, em uma aplicação web utilizada por 51 participantes de diferentes níveis de escolaridade e área de conhecimento.

Os resultados obtidos são promissores e apontam um caminho consistente para a criação de modelos computacionais que utilizam o senso comum como técnica auxiliar formal para aferição de relevância. Tais estudos podem servir de base o desenvolvimento de buscas mais inteligentes de documentos na *Web*. No contexto da sustentabilidade, tais mecanismos podem potencialmente apoiar a definição de políticas públicas de desenvolvimento sustentável, auxiliam gestores na identificação e seleção de projetos que melhor atendam as prioridades definidas, sejam ambientais, sociais e/ou econômicas.

Como trabalhos futuros, pretende-se ampliar os experimentos em dois sentidos. Pela investigação de outras ferramentas de mineração de texto, que podem incluir novos experimentos com a busca por similaridade semântica entre projetos sustentáveis distintos. Em outra direção pretende-se ampliar estudos sobre *crowdsourcing* e formalização da participação humana em mecanismos híbridos de aferição de relevância.

REFERÊNCIAS

PORTO, Vania Christina Nascimento. Bicultivo de Alface e Rúcula Consorciadas com Cenoura em Faixas. 102f. Tese (Doutorado em Fitotecnia) - Universidade Federal Rural do Semi-Árido, Mossoró, 2008

YUEN, Man-Ching; KING, Irwin; LEUNG, Kwong-Sak. A survey of crowdsourcing systems. In: Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom). IEEE, 2011. p. 766-773.

BRACHMAN, Ronald; LEVESQUE, Hector. Knowledge representation and reasoning. Elsevier, 2004.

KOSALA, Raymond; BLOCKEEL, Hendrik. Web mining research: A survey. ACM Sigkdd Explorations Newsletter, v. 2, n. 1, p. 1-15, 2000.

GUPTA, Vishal; LEHAL, Gurpreet S. A Survey of Text Mining Techniques and Applications. 2009.

STUDER, Rudi; BENJAMINS, V. Richard; FENSEL, Dieter. Knowledge engineering: principles and methods. Data & knowledge engineering, v. 25, n. 1, p. 161-197, 1998.

VILLA, Ferdinando; ATHANASIADIS, Ioannis N.; RIZZOLI, Andrea Emilio. Modelling with knowledge: A review of emerging semantic approaches to environmental modelling. Environmental Modelling & Software, v. 24, p. 577-587, 2009.

WIMALASURIYA, Daya C.; DOU, Dejing. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 2010.

LOH, Stanley; DE OLIVEIRA, Jose Palazzo M.; GASTAL, Fábio Leite. Knowledge discovery in textual documentation: qualitative and quantitative analyses. Journal of

Documentation, v. 57, n. 5, p. 577-590, 2001.

FAYYAD, Usama M., et al. Advances in knowledge discovery and data mining. (1996)

SRIVASTAVA, Ashok N.; SAHAMI, Mehran (Ed.). Text mining: classification, clustering, and applications. CRC Press, 2010.

SARASUA, Cristina; SIMPERL, Elena; NOY, Natalya F. Crowdmap: Crowdsourcing ontology alignment with microtasks. In: The Semantic Web–ISWC 2012. Springer Berlin Heidelberg, 2012. p. 525-54.