USO DE VISUALIZAÇÃO E MINERAÇÃO DE TEXTO NO PROCESSO DE ANÁLISE QUALITATIVA: UM ESTUDO DE VIABILIDADE

USING VISUALIZATION AND TEXT MINING TO IMPROVE QUALITATIVE ANALYSIS: A FEASIBILITY STUDY

Elis Cristina Montoro Hernandes
Universidade Federal de São Carlos - UFSCar /Brazil
elis_hernandes@dc.ufscar.br

Emanuel Teodoro
Universidade Federal de São Carlos - UFSCar /Brazil
ema.teodoro@gmail.com

Andre Di Thommazo
Universidade Federal de São Carlos - UFSCar /Brazil
andredt@ifsp.edu.br

Sandra Fabbri
Universidade Federal de São Carlos - UFSCar /Brazil
sfabbri@dc.ufscar.br

***Abstract*: Context: Qualitative analysis is a scientific way to deeply understand qualitative data and to aid in its analysis. However, qualitative analysis is a laborious, time-consuming and subjective process. Aim: The authors propose the use of visualization and text mining to improve the qualitative analysis process. The objective of this paper is to explain how the use of visualization can support the Coding in multiple documents simultaneously, which may allow codes standardization thus making the process more efficient. Method: The Insight tool is being developed to make the proposal feasible and a feasibility study was performed to verify if the proposal offers benefits to the process and improves its results. Results: The study shows that the subjects who applied the proposal got more standardized codes and were more efficient than the ones who applied the process manually. Conclusions: Although the study was conducted in a restrict context (small subject's group and generic data), the results derived from the use of the proposal encourage proceeding with the project.**

***Keywords:* Qualitative analysis; Content analysis, Coding; Feasibility study; Visualization; Text mining; Experimental software engineering; Empirical software engineering.**

***Resumo:* Contexto: análise qualitativa é um método científico que permite compreender detalhadamente dados qualitativos ajudando, portanto, sua análise. Esse tipo de análise é trabalhosa, demorada e subjetiva, e, em geral é conduzida pela técnica de codificação. Objetivo: O objetivo deste artigo é expor como o uso de visualização e mineração de texto pode auxiliar na codificação de diversos documentos simultaneamente, o que pode facilitar a padronização da codificação e, consequentemente, tonar a condução do processo de análise dos dados mais eficiente. Método: para viabilizar o uso de visualização e mineração de texto no processo de codificação, esses recursos estão sendo implementados na ferramenta Insight. Para avaliar a proposta, um estudo de viabilidade foi conduzido para verificar se a proposta oferece benefícios durante o processo e melhora o resultado da análise. Resultados: os resultados indicam que os participantes que utilizaram a proposta criaram códigos mais padronizados e foram mais eficientes se comparados com os participantes que conduziram o processo sem a utilização de visualização e mineração de texto. Conclusão: embora o estudo tenha sido conduzido num contexto limitado (poucos participantes e um domínio de dados genérico), os resultados obtidos encorajam a evolução da proposta e da ferramenta que a torna viável.**

***Palavras-chave:* Análise qualitativa; Análise de conteúdo; Coding; Visualização; Mineração de texto; Estudo de viabilidade; Engenharia de software experimental.**

I. INTRODUÇÃO

De acordo com Basili [1], "*a única maneira de descobrir quão aplicável é um novo método, técnica ou ferramenta em um ambiente específico é experimentar seu uso em tal ambiente*". Considerando essa afirmação, uma das etapas essenciais de uma pesquisa científica é sua avaliação, a fim de obter resultados e definir claramente suas contribuições e, principalmente, suas limitações.

A avaliação de uma pesquisa científica também deve ser feita por meio de um método científico. Em outras palavras, se devem adotar procedimentos bem

definidos para planejamento, coleta e análise dos dados. Em geral, a escolha de como conduzir esses procedimentos está relacionada aos diferentes tipos de estudos experimentais, como exemplo *surveys*, estudo de caso e estudo controlado [2].

Para a etapa de análise dos dados, tanto análise qualitativa quanto quantitativa podem ser aplicadas. A maior diferença entre esses tipos de análise são os tipos de dados a serem analisados – usualmente texto para análise qualitativa e números para análise quantitativa – e o procedimento para obter as conclusões do estudo.

Considerando análise qualitativa, o procedimento básico para análise dos dados se inicia com a identificação de trechos relevantes dos dados coletados.

Após serem identificados, esses trechos relevantes, denominados *quotations*, recebem marcações para que possam ser posteriormente identificados e analisados. Essas marcações, denominadas *labels* ou *codes*, podem ser uma palavra ou uma frase. A proposta nesse momento é identificar diferenças e similaridades nos dados qualitativos coletados.

O objetivo dos *codes* é facilitar a identificação e interpretação de trechos relevantes do texto [3]. Para facilitar a interpretação, os *codes* devem ser organizados em subcategorias e/ou categorias, com o intuito de agrupá-los e organizá-los de maneira coesa.

Após esse processo, chamado de *Coding*, novas informações ou informações complementares sobre o objeto do estudo podem ser identificados.

Neste artigo os nomes utilizados serão *quotations* para os trechos relevantes do texto, *codes* para os códigos/*labels* e *coding*, que é a técnica de análise.

Embora possa parecer simples, [4][5] menciona que "*algumas vezes análise qualitativa é enfadonha, geralmente é tediosa e sempre consome mais tempo que o esperado*".

Quando o volume de dados é grande, o processo de análise qualitativa pode ser mais lento e tedioso. Isso pode induzir ao relaxamento do critério para definição e aplicação dos *codes*, perda de trechos de texto relevantes para o estudo, ou definição de diferentes *codes* para trechos de texto semelhantes, o que pode afetar as conclusões sobre os dados e, consequentemente, as conclusões sobre o estudo conduzido.

Embora análise qualitativa seja um método de análise usualmente aplicado por pesquisadores das áreas médicas e humanidades, recentemente pesquisadores da área de tecnologia vêm adotando esse método de análise em suas pesquisas.

De acordo com Seaman [5], o uso de análise qualitativa vem se intensificando na área de engenharia de software, pois esta é uma área na qual o comportamento humano pode influenciar o uso dos métodos e técnicas. A autora menciona que uma das vantagens em utilizar análise qualitativa é que o pesquisador tende a se aprofundar na complexidade dos dados e nas questões que emergem do estudo e não necessariamente fica concentrado em abstrair informações apenas sobre o que está sendo investigado. Por outro lado, esse tipo de análise exige mais esforço quando comparado aos métodos de análise quantitativa.

Considerando as vantagens em aplicar análise qualitativa na área de engenharia de software e quão difícil é aplicar esse método quando o estudo envolve um grande volume de dados, a proposta aqui apresentada é de utilizar técnicas de visualização e mineração de texto para dar suporte a este processo de análise.

Espera-se coletar evidência de que o uso dessas técnicas torna o processo menos árduo, permitindo que o conjunto de dados a serem analisados (na maioria dos casos, um conjunto de documentos textuais) possa ser manuseado simultaneamente, proporcionando resultados de análises mais consistentes, isto é, *quotations*, *codes* e categorias mais concisas quando comparadas com resultados obtidos sem o suporte dessas técnicas.

Este artigo está organizado em cinco sessões: a Seção 2 apresenta, resumidamente, o conceito de análise qualitativa; a Seção 3 apresenta a proposta, nomeada *Visual Coding*; a Seção 4 relata o estudo de viabilidade conduzido para avaliá-la. Finalmente, a Seção 5 apresenta as conclusões e os estudos em andamento.

II. ANÁLISE QUALITATIVA

De acordo com Strauss and Corbin [6], o objetivo da pesquisa qualitativa é entender um tema específico por meio de descrições, comparações e interpretação de dados, diferentemente da pesquisa quantitativa, que utiliza números para entender determinado tema. Assim, a pesquisa qualitativa diz respeito a um tipo de pesquisa em que os resultados

não são alcançados por meio de procedimentos estatísticos, uma vez que os dados são representados por palavras, imagens, vídeos e sons, e não apenas números.

Coleman e O'Connor [7] comentam que enquanto estudos quantitativos se concentram em questões como "*Quanto?*" e "*Qual a frequência?*", estudos qualitativos estão relacionados a questões tais como "*Por quê*", "*Como?*", "*De qual forma?*". Vale ressaltar que os autores destacam que métodos qualitativos e quantitativos são complementares e, se conduzidos em conjunto, podem melhorar os resultados de uma pesquisa.

Como exemplo de aplicação de análise qualitativa em engenharia de software, é possível considerar um cenário no qual duas técnicas similares estão sendo comparadas e conhecer as razões do porque uma técnica é mais efetiva que a outra pode ser mais relevante que conhecer apenas qual técnica é mais efetiva.

Seaman [4] apresenta métodos de análise qualitativa divididos em dois conjuntos:

- Geração de teoria: são métodos usados para geração de hipóteses fundamentadas nos dados. Exemplos: Método de Comparação Constante e Método de Análise de Casos Cruzados (*Cross-Case Analysis*);
- Confirmação de teoria: são métodos usados para construir uma "evidência de peso" necessária para confirmar alguma hipótese. Exemplos: Validação, Triangulação, Anomalias dos dados, Análise do Caso Negativo e Replicação.

Embora alguns desses métodos possam considerar análise de dados quantitativos, para analisar dados qualitativos (usualmente dados textuais), a técnica *Coding*, brevemente mencionada na seção anterior, é comumente aplicada. Essa técnica é dividida em três passos [5], aqui apresentados com a nomenclatura em inglês:

(i) *Open coding*: o pesquisador deve ler todos os dados procurando por referências sobre o tópico de interesse da pesquisa e deve inserir marcações (*codes*) para cada trecho relevante (*quotation*);

(ii) *Axial coding*: o pesquisador deve organizar os *codes* criando categorias com o intuito de entender melhor os dados analisados;

(iii) *Selective coding*: o pesquisador deve reanalisar os *codes* e categorias e elaborar uma descrição que sintetize os dados analisados.

Hancock [3] descreve a técnica *Coding* por meio de um conjunto de dez passos:

1) Ler os dados textuais procurando por trechos com informações relevantes e escrever uma pequena nota (*code*) que represente essa informação;
2) Elaborar uma lista dos diferentes *codes*;
3) Agrupar os *codes* em categorias, as quais devem representar o tópico principal relacionado aos *codes* agrupados, e elaborar uma lista dessas categorias;
4) Se houver categorias inter-relacionadas, criar uma nova categoria e definir a hierarquia entre essas categorias;
5) Analisar e comparar todas as categorias, mudando suas posições na hierarquia ou criando novas categorias se necessário;
6) Repetir os passos de 1 a 5 para todos os documentos da pesquisa;
7) Certificar-se de que todos os trechos de texto marcados com o mesmo *code* são relacionados;
8) Certificar-se de que as categorias, seus nomes e as hierarquias são representativos;
9) Analisar possíveis relações entre as categorias. Essas relações podem sugerir importantes percepções sobre o estudo/pesquisa. Essa análise deve ser conduzida depois de certificar-se que todos os *codes* e *quotations* estão nas categorias adequadas;
10) Revisar os documentos considerando as categorias criadas e observando trechos de texto não considerados antes, mas que podem se tonar relevantes nesse último momento.

Como esses dez passos sugerem, a técnica de *Coding* é uma tarefa simples, porém árdua, e requer comprometimento e habilidade do pesquisador que vai analisar os dados. Quando a técnica *Coding* deve ser aplicada em um grande volume de dados, alguns

problemas podem ocorrer e dificultar ou prejudicar os resultados obtidos:

- O processo de análise pode ser suscetível ao relaxamento do critério de codificação, ou seja, o pesquisador pode iniciar a análise de forma mais cuidadosa, procurando minuciosamente por trechos relevantes e por detalhes implícitos no texto e, após um período de análise, o mesmo pesquisador pode se tornar menos detalhista. Dessa forma, trechos relevantes podem ser perdidos no decorrer da análise;
- Diferentes *codes* podem ser aplicados a *quotations* similares, uma vez que os dados podem ser analisados em diferentes momentos. Nesse caso, os passos 7 e 8 descritos acima podem exigir esforço adicional.

Algumas ferramentas (software) podem auxiliar a condução de análise qualitativa. Como exemplo, podem-se citar as ferramentas NVivo [8], Atlas.ti [9], The Ethnograph [10], webQDA [`11] e SaturateApp [12]. Ainda assim, alguns pesquisadores relatam o uso de planilhas eletrônicas [13] e software de processamento de texto [4] para auxiliar a condução da análise qualitativa.

Essas ferramentas oferecem muitos recursos para auxiliar a condução do processo de análise qualitativa, sendo que algumas são gratuitas e outras não. No entanto, elas não proveem recursos computacionais para facilitar a análise simultânea de vários documentos em um determinado momento. Embora nessas feramentas os *codes* possam ser reutilizados nos diferentes documentos de um mesmo projeto, a procura por trechos de texto semelhantes deve ser feita pelo próprio pesquisador, o que não evita que os problemas mencionados previamente ocorram.

III. A PROPOSTA *VISUAL CODING*

Como mencionado nas seções anteriores, a análise qualitativa permite explorar questões de maneira detalhada, permitindo aos pesquisadores a obtenção de resultados e conclusões mais relevantes acerca da questão de pesquisa.

Considerando o que foi exposto na seção anterior, a proposta aqui apresentada tem a intenção de tornar o processo de análise qualitativa, particularmente a condução da técnica *Coding*, mais eficiente (rápido) e mais efetivo (melhores resultados).

Apesar do fato da análise qualitativa também ser aplicável em imagens e vídeos, no contexto deste trabalho, consideram-se documentos textuais. Conforme os *codes* são inseridos nos documentos o objetivo do pesquisador é encontrar padrões, de forma que os dados sejam agrupados de acordo com esses padrões. O objetivo é entender detalhadamente os dados e descobrir novas informações de maneira mais fácil.

Assim, a proposta aqui apresentada baseia-se em dois recursos:

i) Visualização, para permitir a navegação em vários documentos de forma simultânea a fim de lidar conjuntamente com informações semelhantes encontradas nos diversos documentos, e

ii) Mineração de texto, para facilitar a busca e identificação de padrões em diferentes documentos [14].

Salienta-se que a premissa é que processar vários documentos ao mesmo tempo, por meio de visualização e mineração de texto, pode ajudar a padronizar a atribuição de *codes* bem como tornar a aplicação da técnica *Coding* mais eficiente.

Para tornar a proposta viável de ser aplicada com base nos recursos mencionados, se fez necessária a implementação de uma ferramenta que foi nomeada *Insight* e que tem sido aprimorada constantemente. A Figura 1 exibe a tela principal dessa ferramenta, na qual algumas partes estão destacadas.

Para apresentar a proposta por meio das funcionalidades disponíveis na versão atual da ferramenta *Insight*, considera-se que há um conjunto de documentos textuais que precisam ser analisados. Após criar um novo projeto na ferramenta, identificando-o por meio de diretório, nome, pesquisador, descrição e idioma, os documentos a serem analisados são inseridos no projeto.

Inicialmente, cada caixa da visualização *TreeMap* [15] representa um documento. Na área de configuração da *TreeMap* (Figura 1 - área C), o pesquisador/usuário da ferramenta pode modificar as cores, texto, tamanho e hierarquia dos dados na visualização, configurando-a da melhor maneira

para atender suas necessidades e facilitar a análise.

Atualmente, a *Insight* utiliza a solução *TreeMap* [16], que é livre, integrada às suas funcionalidades

Para analisar e codificar os documentos usando os recursos atualmente disponíveis na ferramenta *Insight*, o pesquisador deve seguir os seguintes passos:

- Identificar e selecionar um trecho relevante em um dos documentos e criar um código para marcá-lo. Se for apropriado, a opção "*Apply this code for equal quotations*" (aplicar o *code* para *quotations* iguais) irá aplicar o mesmo código em trechos de texto idênticos ao que foi selecionado, nos diversos documentos;
- Inserir uma palavra-chave do trecho selecionado ou inserir o trecho completo na área de busca (Figura 1 - área D). Se uma ou mais caixas da visualização *TreeMap* (Figura 1 - área B) forem destacadas (receberem a borda azul), significa que o documento representado pela caixa possui o trecho ou a palavra-chave procurada;
- Clicando em uma das caixas destacadas o documento correspondente será aberto na área de codificação (Figura 1 - área A) e a palavra-chave ou o termo será destacado na tela;
- Depois de ler o trecho destacado, é possível aplicar o mesmo *code* recentemente aplicado, definir um novo *code* ou reutilizar um *code* aplicado em outro momento.
- Se mais de uma caixa (ou seja, documento) estiver destacado na visualização, o pesquisador pode abrir, analisar e codificar todos eles sequencialmente. A possibilidade de codificar trechos idênticos ou com a mesma palavra-chave em momentos próximos pode facilitar a interpretação dos dados e a definição dos *codes*, o que facilita a padronização da codificação;

Outra funcionalidade disponível é a possibilidade de analisar os dados, criar e aplicar *codes* baseando-se em similaridade dos dados. Assim, selecionando-se um trecho relevante e escolhendo-se a opção "*Mining this quotation*" (minere essa *quotation*), uma nova *TreeMap* é exibida. As caixas dessa nova *TreeMap* representam os diversos documentos sob análise, e a cor dessas caixas indica a porcentagem de similaridade que cada documento possui em relação ao trecho selecionado. Uma legenda exibida na tela associa a cor com a referida porcentagem. A Figura 3 exibe um exemplo dessa funcionalidade.

O cálculo de similaridade entre o trecho selecionado pelo pesquisador e todos os outros documentos é feito por meio dos métodos *Frequency Vector* e *Cosine Similarity* [17].

Para dar suporte às etapas de *Axial Coding* e *Selective Coding* mencionados na Seção 2, a ferramenta *Insight* permite agrupar os *codes* em categorias, e as categorias em outras categorias, sem restrição de níveis. A Figura 4 exibe um exemplo dessa funcionalidade.

Um editor de texto também é disponibilizado pela ferramenta para permitir que o pesquisador registre seus comentários, ideias, hipóteses e até mesmo teorias que emergiram durante a análise.

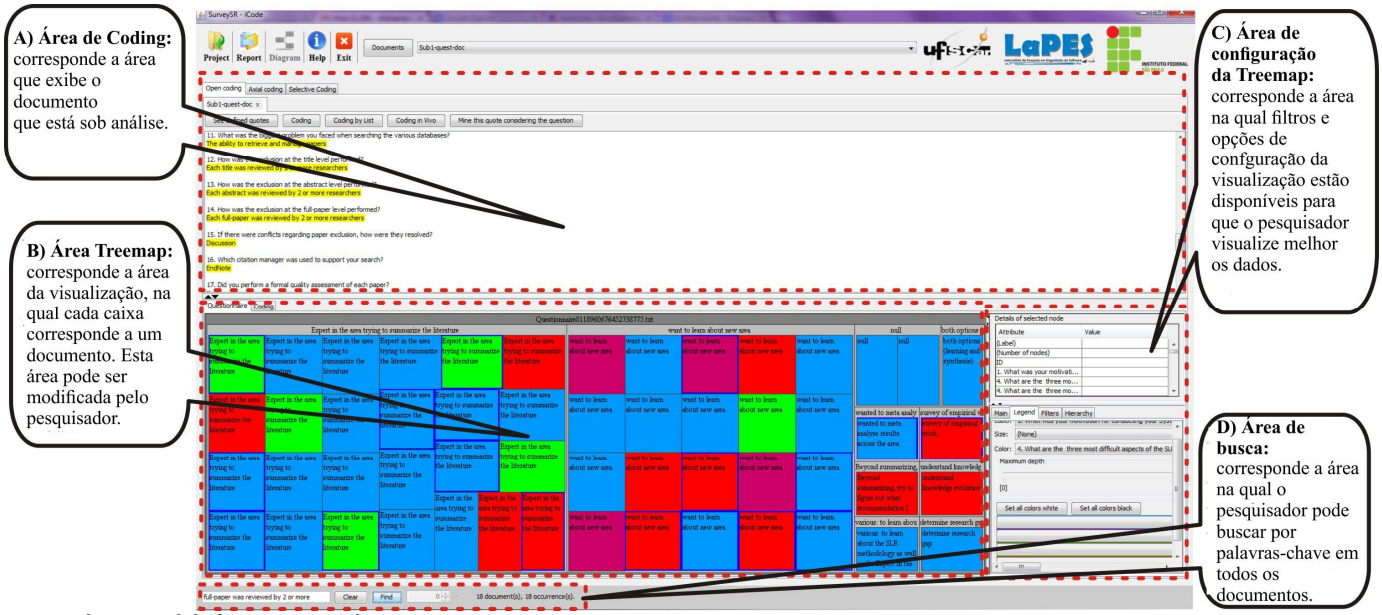

Fig. 1. Tela principal da ferramenta *Insight*.

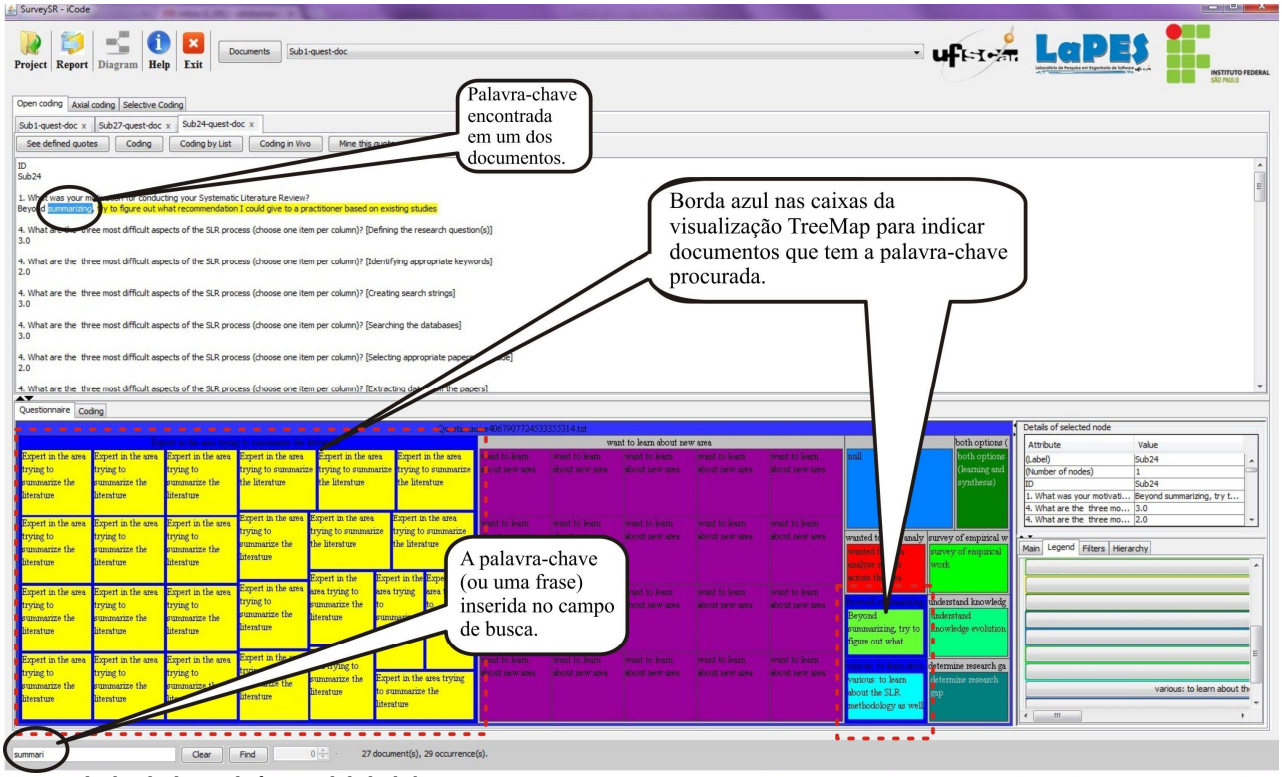

Fig. 2. Exemplo de tela do uso da funcionalidade de busca.

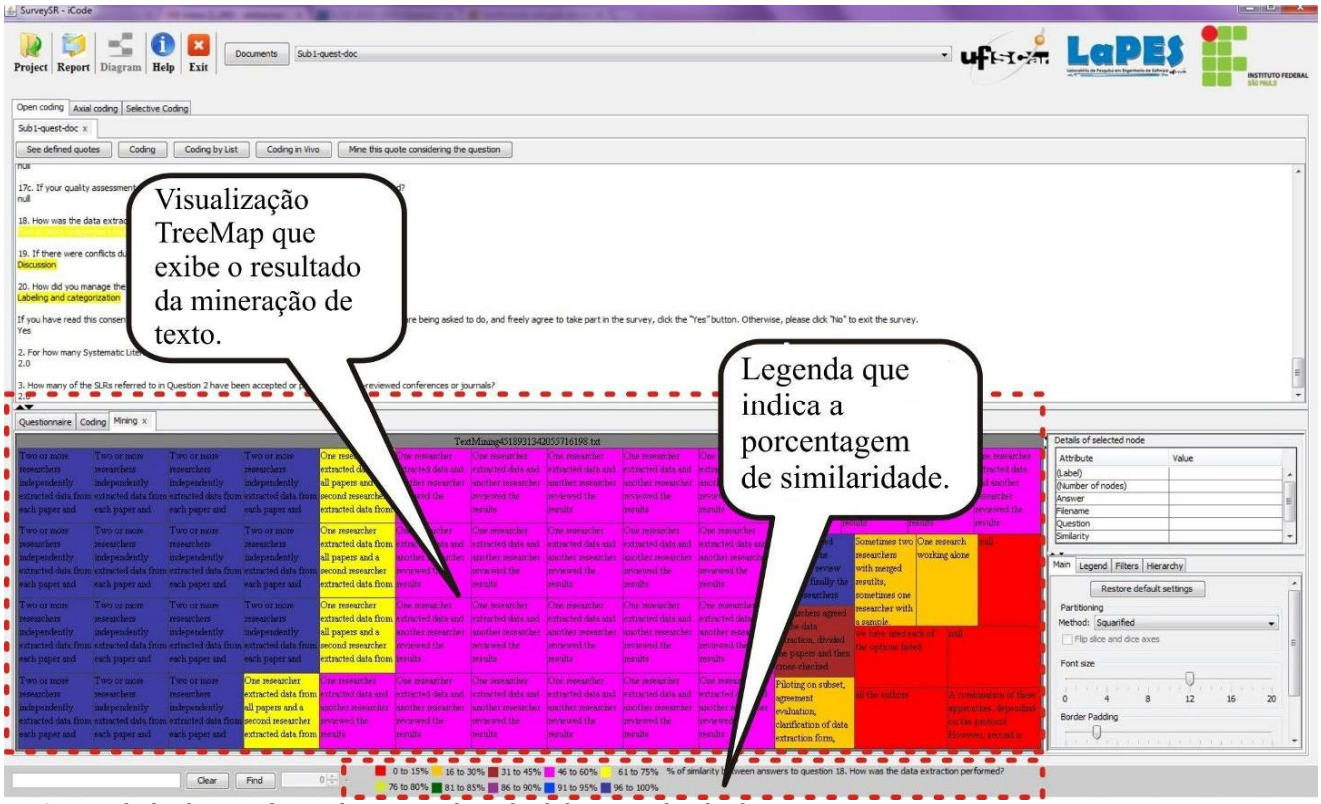

Fig. 3. Exemplo de tela para aplicar *codes* por meio de similaridade entre trechos dos documentos.

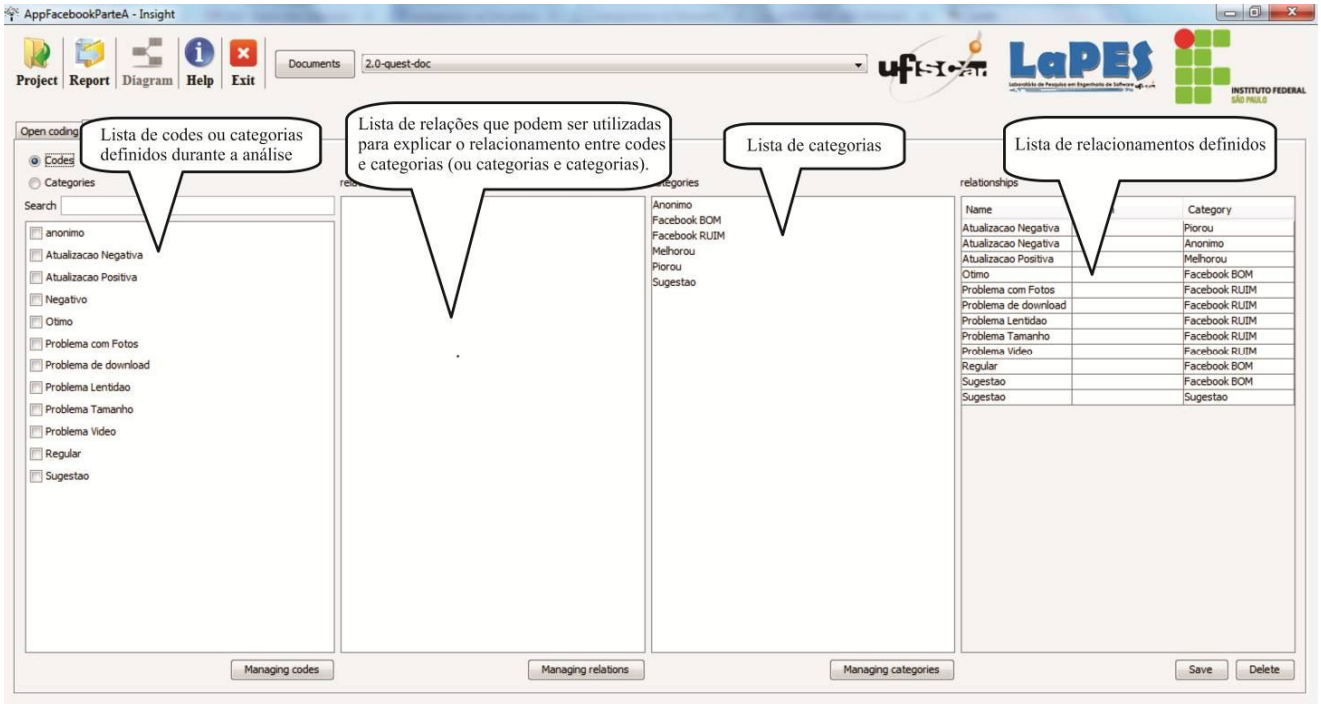

Fig. 4. Exemplo de tela para *Axial Coding*.

IV. ESTUDO DE VIABILIDADE

De acordo com a metodologia experimental para introduzir processos de *software* apresentada em [18], foi conduzido um estudo de viabilidade para avaliar a proposta de utilizar visualização e mineração de texto para facilitar a análise qualitativa, especificamente a técnica *Coding*.

O processo tradicional de *Coding* foi modificado por meio da inserção da visualização e mineração de texto, tornando mais fácil a análise e codificação de um conjunto de documentos, por meio do tratamento simultâneo de vários documentos em um momento de interesse. O objetivo do estudo foi coletar evidências de que essa forma de manusear os documentos torna a condução da análise qualitativa mais eficiente e efetiva.

Nas próximas subseções são apresentadas as etapas do estudo: a Subseção A apresenta a identificação, definição e planejamento do estudo; a Subseção B apresenta a condução do estudo; a Subseção C apresenta a análise dos dados, resultados e discussões e, por fim, a Subseção D apresenta as ameaças à validade do estudo.

A. *Identificação, definição e planejamento*

O estudo foi planejado utilizando o modelo GQM [18], conforme apresentado na Tabela I.

Para esse estudo foram criados nove artefatos:

(i) material de treinamento sobre análise qualitativa e a técnica *Coding*,
(ii) material de treinamento sobre a técnica de visualização *TreeMap*;
(iii) material de treinamento sobre a ferramenta *Insight, uma vez que por ela é que se consegue aplicar visualização e mineração de texto*;
(iv) questionário de caracterização dos participantes;
(v) formulário de consentimento para os participantes;
(vi) dados para a aplicação do *Coding* - seleção de reportagens jornalísticas sobre a Copa do Mundo de Futebol de 2014;
(vii) formulário para reportar o resultado;
(viii) questionário de *feedback*;
(ix) modelo de referência da aplicação da técnica *Coding* nas reportagens selecionadas.

TABELA I. OBJETIVO DO ESTUDO DE VIABILIDADE

Analisar	O uso de visualização e mineração de texto para conduzir a técnica *Coding*
Com o propósito de	Avaliar a viabilidade
Com respeito a	Efetividade (padronização dos *codes*) e eficiência (tempo gasto)
Do ponto de vista dos	Pesquisadores
No contexto de	Alunos de mestrado e doutorado

É importante mencionar que o modelo de referência foi elaborado manualmente por um dos autores deste artigo e revisador por outro autor, e utilizado como parâmetro para comparação do resultado dos participantes. Salienta-se que esse modelo não é considerado a versão correta da análise, mas sim como uma versão criada por pessoas com maior experiência na técnica *Coding* e em análise qualitativa.

O desenho do estudo foi definido com o intuito de identificar os efeitos em conduzir a técnica *Coding* em um conjunto de documentos por meio de diferentes procedimentos:

- explorando o uso da visualização e mineração de texto, por meio da ferramenta *Insight*, de forma a poder processar a codificação em vários documentos em um determinado momento;
- analisando os documentos separadamente, por meio da ferramenta *Insight*, da forma como a codificação é comumente realizada quando se usam outras ferramentas;
- analisando os documentos separadamente, de forma manual, como geralmente é feito quando não se tem acesso a ferramentas.

Para selecionar os participantes, uma mensagem foi enviada em uma lista de e-mails de estudantes de pós-graduação. Um mestrando e cinco doutorandos participaram do estudo como voluntários. Após a aplicação do questionário de caracterização, os participantes foram divididos em grupos, de acordo com suas habilidades, resumidas na Tabela II.

TABELA II. CARACTERIZAÇÃO DOS PARTICIPANTES.

Questão	Amplo conhecimento	Pouco conhecimento	Nenhum conhecimento
1) Qual o seu nível de conhecimento teórico sobre análise qualitativa?	P1, P2, P4		P1, P5, P6
2) Qual o seu nível de conhecimento sobre a técnica *Coding*	P5, P6	P1, P2, P3, P4	
3) Qual o seu nível de conhecimento sobre a técnica de visualização *TreeMap*?	P5, P6	P1, P3	P2, P4
4) Qual o seu nível de conhecimento sobre a ferramenta *TreeMap*?	P5,	P1, P3, P6	P2, P4

Com base no resultado do questionário de caracterização os seis participantes, nomeados como P1, P2, P3, P4, P5 e P6, foram distribuídos em três grupos, para contemplar os três procedimentos diferentes. A Tabela III sumariza essa informação sendo que o Grupo C foi o grupo de controle, uma vez que essa opção é a comumente utilizada. Os grupos A e B foram os grupos de tratamento, já que eram essas as opções que estavam sendo avaliadas no estudo.

TABELA III. Definição dos grupos do estudo.

Grupo	**A**	**B**	**C**
	Tratamento	Tratamento	Controle
Participantes	(P1 & P2)	(P3 & P4)	(P5 & P6).
Forma de conduzir a técnica *Coding*	(ferramenta *Insight* + análise simultânea)	(ferramenta *Insight* + análise isolada)	(manualmente + análise isolada)

É importante ressaltar que os participantes do Grupo C puderam manusear os arquivos por meio de software comum de pacotes de escritório, mas também receberam uma cópia impressa das reportagens. Outro motivo que levou a definir as condições do grupo de controle é que embora existam diversos *software* para dar suporte à análise qualitativa, essas ferramentas são proprietárias e suas versões demo apresentam limitações que prejudicariam o estudo em questão.

B. Condução do estudo

O estudo foi conduzido em dois dias. No primeiro dia, na parte da manhã, um instrutor (um dos autores) apresentou a proposta do estudo e obteve o consentimento dos participantes. Os participantes também responderam ao questionário de caracterização. Feito isso, os participantes receberam treinamento sobre análise qualitativa e a técnica *Coding*. Ainda no primeiro dia, na parte da tarde, os participantes dos grupos A e B receberam treinamento sobre a ferramenta *Insight* e sobre a *TreeMap* para aprenderem como a ferramenta *Insight* funciona e como a visualização *TreeMap* poderia ser utilizada durante a análise.

No segundo dia todos os participantes aplicaram a técnica *Coding*. Os documentos analisados foram reportagens jornalísticas sobre a Copa do Mundo de Futebol de 2014. O objetivo foi identificar tópicos de destaque sobre o tema nos diversos documentos. Os participantes conduziram a atividade no mesmo local e nenhum tipo de comunicação foi permitida.

C. Análise dos dados, resultados e discussões

Com o intuito de verificar a viabilidade em conduzir a técnica *Coding* manipulando vários documentos simultaneamente, os dados coletados durante o estudo foram analisados com base no significado dos *codes*; no número de categorias, *codes* e *quotations*; no tempo despendido pelo grupo de tratamento; e, finalmente, na análise qualitativa do questionário de *feedback* respondido pelos participantes.

Além disso, a razão entre o número de *quotations* dividido pelo número de *codes* foi medida com o intuito de observar a consistência dos *codes* definidos por cada participante. Em outras palavras, essa razão significa, neste estudo, quão reutilizado um *code* foi durante a análise.

1) *Análise dos resultados dos participantes*

Considerando que no contexto de análise qualitativa não é viável discutir se um resultado é errado ou melhor que outro, a análise dos resultados deste estudo foi realizada do ponto de vista semântico. Os dados foram analisados seguindo os seguintes passos:

i) os *codes* de cada participante e os *codes* do modelo de referência foram tabulados;

ii) baseando-se na semântica (no significado dos codes), os *codes* de cada participante foram comparados com os *codes* do modelo de referência, com o objetivo de identificar *codes* relacionados;

iii) *codes* relacionados foram destacados.

A Tabela 4 apresenta a sumarização dos resultados. Observa-se que o número de *quotations* relacionadas a cada *code* não é exibida. Dados complementares como o número de categorias, *codes* e *quotations* são apresentados na Figura 3.

É possível observar que todos os *codes* definidos no modelo de referência foram definidos por, ao menos, um dos outros participantes. Além disso, cinco dos seis participantes apresentaram 50% ou mais de *codes* semanticamente equivalentes aos *codes* do modelo de referência. Esse fato sugere que os participantes entenderam o objetivo do estudo, assim como a técnica *Coding*.

Em relação à efetividade, ou seja, a padronização dos *codes*, o Grupo A apresentou resultados mais padronizados e homegêneos se comparados com o

resultado dos outros participantes, uma vez que o número de *codes* definidos por esses participantes (19 e 15) foram os menores, e a razão *"nº. de quotations dividido por nº. de codes"* de ambos os participantes foram similares (2,1 e 2,2). Esse resultado sugere que codificar (*Coding*) um conjunto de documentos tendo a possibilidade de tratar vários documentos ao mesmo tempo (simultaneamente) pode facilitar o reuso dos *codes*, tornando-os mais consistentes.

O Grupo B não apresentou resultados homogêneos. As razões *"nº. de quotations dividido por nº. de codes"* dos participantes P3 e P4 foram distintas (3,3 e 1,6). Esse resultado sugere que a efetividade do participante P3 está mais relacionada à habilidade pessoal que ao procedimento utilizado para analisar os dados.

Embora o participante P3 tenha conduzido o *Coding* por meio da ferramenta *Insight* e, consequentemente, utilizado a visualização, o que deveria ter facilitado o reuso de *codes*, esse participante declarou no questionário de feedback que "*a visualização TreeMap ajudou, mas é necessário treinar mais o uso dessa técnica para usufruir dos benefícios que ela pode oferecer no Coding*". Por outro lado, o participante P4, que criou poucas categorias, comentou que "*a ferramenta Insight deveria prover uma maneira mais fácil de criar as categorias, como por exemplo, uma funcionalidade de drag and drop*".

TABELA IV. SUMARIZAÇÃO DOS RESULTADOS DOS PARTICIPANTES

***Codes* do modelo de referência**	**Grupo A**		**Grupo B**		**Grupo C**	
	P1	**P2**	**P3**	**P4**	**P5**	**P6**
Investimento em aeroportos	X	X	X		X	X
Privatização de aeroportos	X	X		X		
Problemas em aeroportos	X	X	X		X	X
Benefício em impostos		X		X		X
Benefícios para pessoas com necessidades especiais	X	X	X	X		
Benefícios econômicos	X		X	X		
Geração de empregos	X	X			X	
Voluntariado					X	
Benefícios em infraestrutura			X	X	X	X
Ingressos de baixo custo	X				X	X
Valor dos ingressos		X	X	X		X
Beneficiários do Bolsa Família"					X	
Bebidas alcoólicas nos estádios	X	X	X	X	X	X
Melhorias e reformas nos estádios	X	X				X
Legislação	X	X	X		X	X
Visto	X	X	X	X	X	X
Recomendações internacionais					X	
Discussão pública sobre o tema					X	
Organização		X				
Segurança	X	X	X		X	
Cronograma	X	X	X			X
turismo		X	X			

Número de *codes* semelhantes	***13***	***15***	***12***	***8***	***13***	***11***
Número de *codes* não semelhantes	***6***	***0***	***12***	***12***	***7***	***6***
Número de *quotations* dividido por número de *codes"*	***2,1***	***2,2***	***3,3***	***1,6***	***1,2***	***1,2***

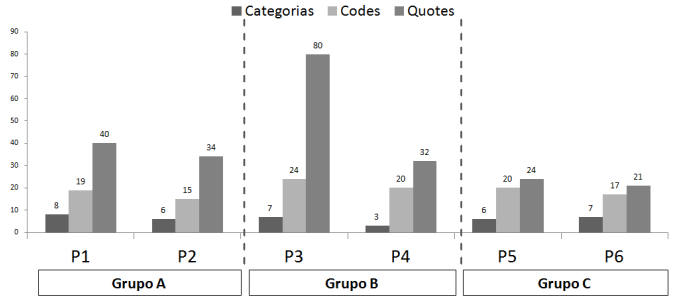

Fig. 5. Número de categorias, codes e quotations de cada participante.

O Grupo C apresentou resultados iguais, mas também apresentou os valores mais baixos para a razão *"nº. de quotations dividido por nº. de codes"* (1,2 e 1,2), uma vez que o número de *quotations* e *codes* foram similares. Esse resultado, em conjunto com feedback dados pelos participantes (apresentado no item 2 dessa subseção) , sugere que conduzir a técnica *Coding* manualmente é trabalhoso e pode ser entediante, o que pode dificultar o reuso de *codes* e a identificação de *quotations.*

Como mencionado anteriormente, o número de *codes* não é um dado crucial para análise qualitativa, ao menos que a técnica *Coding* esteja sendo conduzida para transformar dados qualitativos em dados quantitativos, o que não é o objetivo do estudo apresentado. No entanto, comparando os *codes* definidos pelos participantes P5 e P6 com os do modelo de referência, é possível observar que informações relevantes dos dados analisados não foram identificadas pelos participantes. Esse fato pode prejudicar a sumarização final da análise qualitativa dos dados.

1) Análise do tempo despendido por participante

Considerando que a proposta apresentada neste artigo é no contexto de análise qualitativa e não é possível assumir uma versão oráculo para comparar os resultados da técnica *Coding* gerada por cada participante, o tempo despendido por cada um deles é um bom dado para avaliar a viabilidade da proposta.

Na Figura 8 é apresentado o tempo de análise de cada participante. Os participantes do Grupo A apresentaram os menores tempos para realizar a análise. Esses participantes mencionaram que "*a*

possibilidade de navegar entre os documentos por meio da visualização ajudou a economizar tempo".

Considerando o Grupo B, o participante P3 despendeu mais tempo dentre os participantes que utilizaram a ferramenta Insight. Observando a Figura 5 é possível notar que esse participante é bem detalhista, identificando o maior número de *quotations*. Por outro lado, o participante P4, também do Grupo B, realizou a atividade em menos tempo, mas criou poucas categorias.

O tempo despendido pelo Grupo C foi maior que o tempo dos grupos A e B. Esse resultado era esperado e corrobora o fato de que, em geral, realizar uma atividade de forma manual é mais demorado do que realizar a mesma atividade com o auxílio de alguma ferramenta computacional que dê suporte à realização dessa atividade.

Comparando o menor e o maior tempo despendido de cada grupo na atividade pode-se observar que o Grupo A foi em torno de 55% mais eficiente que o Grupo C e o Grupo B foi em torno de 33% mais eficiente que o Grupo C:

- *Grupo A x C:* (i) maior tempo: P2 e P6. P2 foi 55.56% mais eficiente que P6; (ii) menor tempo: P1 e P5 - P1 foi 55.18% mais eficiente que P5;
- *Group B x C:* (i) maior tempo: P3 e P6. P3 foi 34.49% mais eficiente que P6; (ii) menor tempo: P4 e P5. P4foi 33.34% mais eficiente que P5.

Em resumo, o tempo despendido pelos participantes somados aos resultados apresentados na subseção anterior sugere que analisar vários documentos ao mesmo tempo, à medida que os *codes* são definidos, com a ajuda de visualização e mineração de texto, é uma abordagem promissora, cenário que estimula a continuidade desta pesquisa.

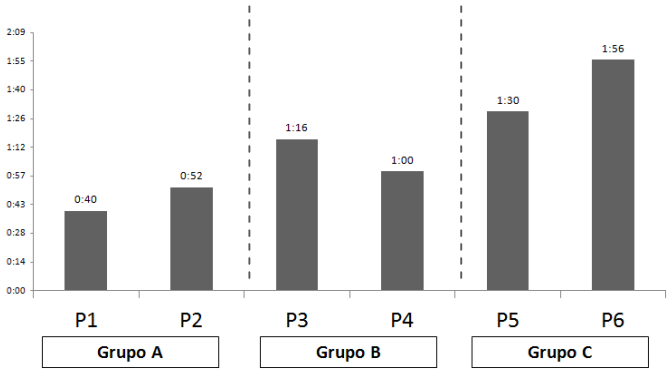

Fig. 6. Tempo de análise de cada participante

2) Análise do questionário de feedback dos participantes

Depois de realizar a atividade, ou seja, aplicarem a técnica *Coding*, cada participante respondeu a um questionário de *feedback* de acordo com o seu grupo. As respostas foram analisadas por meio da técnica *Coding*, com o suporte da ferramenta *Insight*. O resultado é apresentado na Tabela V.

Os participantes do Grupo A responderam as questões 1, 2, 3 e 4; os do grupo B a questão 1 e os do Grupo C a questão 4.

A questão 1 é relacionada ao uso da visualização *TreeMap*. De acordo com os participantes, o uso da visualização é útil para: (i) identificar rapidamente o resultado de uma busca, o que ajuda a gerenciar vários documentos ao mesmo tempo; e (ii) ajuda na execução do *Coding*, permitindo visualizar e reutilizar *codes* que já foram criados em outros momentos e contextos.

Esses benefícios mencionados pelos participantes estão de acordo com a intenção dos autores da proposta. Por outro lado, um dos participantes indicou que o treinamento na *TreeMap* não foi suficiente. Como mencionado na subseção anterior (Análise dos resultados dos participantes), a falta de habilidade em utilizar a técnica de visualização pode ser um fator que interferiu no rendimento do participante. Esse comentário é importante para que os autores melhorem o planejamento do próximo estudo, tanto em tempo de treinamento da *Treemap* quanto em relação ao material de treinamento.

A questão 2 é relacionada à funcionalidade de busca, que combinada com a visualização deve ajudar a análise simultânea dos documentos. De acordo com os participantes, essa funcionalidade foi utilizada com frequência para localizar trechos de

texto iguais e reutilizar os *codes* de forma coerente, promovendo a padronização dos *codes* e *quotations*.

A questão 3 é relacionada à funcionalidade de mineração de texto que, combinada com a visualização, também deve ajudar a análise simultânea dos documentos. De acordo com os dois participantes essa funcionalidade foi utilizada quando uma *quotation* longa era identificada e o participante queria conferir se havia algum trecho de texto semelhante no qual o mesmo *code* poderia ser aplicado.

A questão 4 foi relacionada à análise simultânea - usando a visualização e selecionando a informação que deveria ser apresentada nas caixas da *TreeMap*, ou utilizado a visualização em conjunto com o recurso de busca para encontrar trechos de texto iguais. De acordo com os participantes, a possibilidade de lidar com vários documentos simultaneamente ajudou a ler os documentos e padronizar a aplicação dos *codes* (aplicar o mesmo *code* em trechos semanticamente equivalentes).

As respostas das questões 2, 3 e 4 também fortalecem a intenção dos autores em relação aos benefícios da proposta.

A questão 5 é relacionada às dificuldades encontradas pelos participantes que realizaram a atividade manualmente. De acordo com esses participantes, a condução manual da técnica *Coding* leva a analisar um documento por vez, já que tentar analisar todos em conjunto é uma tarefa difícil. Assim, as dificuldades mencionadas pelos dois participantes foram: (i) a criação de diferentes *codes* para o mesmo assunto e (ii) a dificuldade em definir as categorias.

A intenção dos autores relacionada a essa questão é identificar funcionalidades que devem ser inseridas na ferramenta *Insight*. As funcionalidades requeridas pelos participantes do Grupo C estão relacionadas ao gerenciamento de *codes* e categorias, à ajuda para vasculhar os documentos em busca de informações relevantes e por fim, à ajuda para evitar a definição de *codes* diferentes para o mesmo assunto.

Algumas dessas dificuldades e requisições de funcionalidades estão contempladas na versão atual da ferramenta. Por outro lado, o foco da proposta é a possibilidade de análise de vários documentos ao mesmo tempo, o que foi mencionado como uma necessidade pelos participantes do Grupo C.

Em resumo, a análise do questionário de *feedback* sugere que a funcionalidade de busca combinada com a visualização proporciona facilidades ao *Coding*, o que indica que a proposta é viável e deve ter continuidade.

TABELA V. RESULTADO DA ANÁLISE DO QUESTIONÁRIO DE FEEDBACK

Questões	**Categorias**	***Codes***	**nº. *Quotations***
Q1) Qual a sua opinião sobre o uso da *TreeMap* no contexto da técnica *Coding*? A visualização ajudou na interpretação das informações e na geração de novos *codes*?	Visualização dos resultados da busca	Isso ajudou a visualizar o resultado da busca	2
		TreeMap ajuda a ter uma visão melhor do resultado da visualização	1
	Ajuda na codificação	A visualização ajudou a saber quais documentos deveriam ser lidos	1
		A *TreeMap* ajudou a criar novos *codes*	1
		Ajuda a vasculhar os documentos que possuem determinada palavra	1
	Necessidade de treinamento	Preciso de mais treinamento na *TreeMap*	1
	Visualização dos *codes*	A *TreeMap* ajudou a visualizar todos os codes já criados em cada documento	1
	Reuso dos *codes*	Ajudou a reutilizar *codes*	1
Q2) A funcionalidade de busca foi utilizada com frequência? Por quê?	Usado todo o tempo	A busca foi utilizada o tempo todo	2
	Identificação de *quotations*	Ajuda a vasculhar os documentos que possuem determinada palavra	1
	Padronização dos *codes*	A busca ajuda a aplicar o mesmo *code* em trechos relacionados	1
Q3) A funcionalidade de mineração de texto foi utilizada com frequência? Por quê?	Identificação de *quotations*	A mineração de texto foi utilizada quando uma *quotation* era muito grande e seria interessante buscar trechos similares para aplicar o mesmo *code*	2
Q4) A análise de vários documentos simultaneamente economizou tempo? Por quê?	Ajuda na codificação	Ajuda a vasculhar os documentos que possuem determinada palavra	1
	Padronização dos *codes*	Ajudou a reutilizar *codes*	1

Q5) Na sua opinião, qual foi a maior dificuldade durante a condução da atividade manualmente e o que você espera de uma ferramenta para dar suporte ao *Coding*?	Necessidade de funcionalidades	Não tinha certeza se estava dando certo ler e codificar mais que um documento ao mesmo tempo, então eu decidi ler e codificar um a um	1
		Suporte para vasculhar os documentos de forma	2
		Suporte para gerenciar categorias	1
		Suporte para gerenciar os *codes*	2
		Suporte para evitar a criação de *codes* repetidos ou semelhantes	1
	dificuldades	Fiquei inseguro durante a definição das categorias	1
		Eu não lembrava o significado de um *code*	1
		Criei diferentes *codes* para o mesmo assunto	1
		Definir as *quotations* e os *codes* manualmente é difícil	1

D. *Ameaças à validade*

Ameaças à validade são inerentes a qualquer estudo experimental, independentemente do design do estudo. Sendo assim, de acordo com [20], todo relato de estudo deve expor essas ameaças.

Em relação ao estudo de viabilidade apresentado, é possível identificar ameaças à validade interna, externa e de conclusão, como segue:

- Validade interna: o tópico a ser analisado pelos participantes é considerado uma ameaça, pois cada participante pode ter um conhecimento diferente sobre o assunto. A fim de minimizar essa ameaça, os autores escolheram um tópico sobre um assunto genérico - Copa do Mundo de Futebol de 2014. Embora esse assunto não represente um domínio de pesquisa, essa foi a maneira encontrada pelos pesquisadores para minimizar a diferença de conhecimento dos participantes sobre os documentos que seriam analisados;

- Validade externa: é plausível assumir que os resultados poderiam ser diferentes com um conjunto diferente de participantes. Os participantes do estudo apresentado são estudantes de pós-graduação e a maioria possuía pouco conhecimento sobre análise qualitativa. No entanto, considerando os resultados positivos mesmo nesse grupo pouco experiente, entende-se que a proposta pode ser considerada viável para o contexto no qual se insere.

- Validade de conclusão: um dos desafios desse estudo foi a análise dos resultados do *Coding* de cada participante e a caracterização da efetividade da proposta. Para minimizar os riscos relacionados à análise, os autores usaram um modelo de referência para efetuar a comparação dos resultados. No entanto, essa comparação também pode ser considerada uma ameaça à validade, uma vez que a comparação, por ela mesma, assim como o modelo de referência são subjetivos.

V. CONCLUSÕES E ESTUDOS FUTUROS

Análise qualitativa é relevante para engenharia de software, uma vez que uma das características dessa área é a junção de questões técnicas e não técnicas [5]. O sucesso da condução dos processos da área de engenharia de software depende do processo assim como de quem conduz o processo.

Apesar das vantagens que esse tipo de análise oferece aos pesquisadores, sua condução é trabalhosa, consome muito tempo, é suscetível a erro e requer habilidade para ser conduzida corretamente.

Essas características são ainda mais presentes quando há um grande volume de dados a serem analisados. Somado a isso, para conduzir a análise qualitativa, em geral é utilizada a técnica *Coding*, que usualmente é aplicada em um documento por vez. Esse procedimento torna a condução da técnica *Coding* mais difícil e tende a dificultar a padronização dos *codes* à medida que a análise é realizada.

Considerando esse contexto, para aprimorar a aplicação do *Coding*, este artigo apresentou a proposta de utilizar visualização e mineração de texto para permitir a análise de um conjunto de documentos ao mesmo tempo. Essa forma de

conduzir o *Coding* pode permitir que a técnica seja conduzida de forma mais ágil e padronizada, uma vez que quando uma *quotation* é identificada e um *code* é criado, *quotations* similares podem ser identificadas em todos os outros documentos e o mesmo *code* pode ser aplicado a elas.

Dessa forma, este artigo apresentou o conceito de análise simultânea no contexto de *Coding*, exemplificando os passos que compõem a execução da técnica com o auxílio de visualização e mineração de texto. Para que a proposta se tornasse aplicável, era indispensável que uma ferramenta computacional – *Insight* – fosse desenvolvida, para permitir o uso de visualização e mineração de texto.

Por meio de um estudo experimental a proposta foi apresentada e sua viabilidade foi estudada.

Em relação à efetividade o resultado do estudo mostra que o Grupo A apresentou resultados mais padronizados e homogêneos quando comparados com os outros grupos. Esse resultado sugere que codificar (*Coding*) um conjunto de documentos, tendo a possibilidade de manuseá-los ao mesmo tempo, simultaneamente, pode facilitar o reuso dos *codes*. Em relação à eficiência, o resultado sugere que codificar os dados da forma como a proposta aqui apresentada sugere é mais ágil quando comparada com a condução manual, o que já era esperado.

Com base nos comentários do *feedback* dos participantes foram encontrados indícios de que as funcionalidades de busca e mineração de texto combinadas com a visualização tornam a condução da técnica *Coding* mais fácil, fato que pode influenciar na melhoria do processo de análise qualitativa e seus resultados.

Considerando a experiência obtida na condução desse estudo, é possível mencionar duas considerações. Primeiro, a dificuldade em analisar o resultado dos participantes, já que o contexto de análise qualitativa não é apropriado para estabelecer um oráculo. Um modelo de referência foi criado para comparar os resultados dos participantes a fim de minimizar a hipótese de que os participantes tivessem conduzido a análise de maneira equivocada.

Em segundo lugar, durante o estudo foi solicitado que os participantes identificassem informações relevantes nos documentos analisados. Talvez, estabelecer uma perspectiva comum para a análise dos dados possa tornar os resultados mais precisos e fáceis de serem comparados.

Apesar dessas questões, considera-se que a proposta apresentada é viável e promissora. Como trabalhos futuros, espera-se disponibilizar a ferramenta *Insight* sob uma licença GLP. Além disso, planeja-se explorar a proposta no contexto de estudos experimentais para analisar questionário de *feedback* e outro estudo para comparar o uso da *Insight* com o uso da ferramenta Atlas.ti ou NVivo.

Somado a esses trabalhos futuros, atualmente a proposta está sendo utilizada para analisar estudos primários no contexto de estudos secundários (Síntese Temática) [21] e no contexto do processo de inspeção de *software* para analisar lista de defeitos.

REFERENCES

[1] G. Basili, V. Basili., S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sorumgard, M. Zelkowitz Packaging researcher experience to assist replication of experiments. ISERN Meeting, Sydney, pp. 3-6. 1996.

[2] C. Wohlin, P. Runeson, M. Höst, Experimentation in Software Engineering - An Introduction, Sweden: Springer, 2000.

[3] B. Hancock, Trent Focus for Research and Development in Primary Health Care: An Introduction to Qualitative Research. Trent Focus, 2002.

[4] C. Seaman, Qualitative methods in empirical studies of software engineering, IEEE Transactions on Software Engineering, vol. 25, n. 4, p.557-573, Jul./Aug. 1999.

[5] C. Seaman, "Qualitative Methods" in Guide to Advanced Empirical Software Engineering, F. Shull, J. Singer, D. Sjoberg, Eds. London: Springer, 2008, pp. 35–62.

[6] A. Strauss, J. Corbin, Basics of qualitative research: techniques and procedures for developing grounded theory. Sage Publications, 3ed, 2008.

[7] G. Coleman, T. O'Connor, Using grounded theory to understand software process improvement: A study of Irish software product companies. Information and Software Technology. vol.49, n.6, p. 654-667, Fev. 2007.

[8] NVivo. QSR International. Available at: http://www.qsrinternational.com/products_nvivo.aspx.

[9] Atlas.ti: The Qualitative Data Analysis & Research Software. Available at: http://www.atlasti.com/index.html

[10] The Ethnograph. Qualittaive data analysis software. Available at: http://www.qualisresearch.com/

[11] WebQDA. Available at: http://www.webqda.com

[12] Saturate. Simple collaborative qualitative analysis. Available at: http://www.saturateapp.com/

[13] Q. Gu, P. Lago, Exploring service-oriented system engineering challenges: A systematic literature review. Service Oriented Computing and Applications, vol. 3, n. 3, p.171-188, Sep. 2009.

[14] R, Feldman, J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press, 2007.

[15] B. Johnson, B. Shneiderman, Tree-maps: a space-filling approach to the visualization of hierarchical information structures, IEEE Conference on Visualization, VIS, October 1991, San Diego. pp. 284-291.

[16] B. Bederson, B. Shneiderman, M. Wattenberg. Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies. ACM Transactions on Graphics (TOG), vol. 21 (4),p. 833-854. Oct. 2002.

[17] G. Salton, J. Allan. Text retrieval using the vector processing model. 3rd Symposium on Document Analysis and Information Retrieval, University of Nevada, April. 1994, Las Vegas.

[18] F. Shull, J. Carver and G. Travassos, "An empirical methodology for introducing software processes", European Software Engineering Conference, pp. 288-296, September 2001 [9th ACM SIGSOFT International Symposium on Foundations of Software Engineering (ESEC/FSE)].

[19] V. Basili, C. Caldiera, H. Rombach, "Goal Question Metric Paradigm", in Encyclopedia of Software Engineering, J. J. Marciniak, Eds. London: John Wiley & Sons, 1996.

[20] A. Jedlitschka, M. Ciolkowski and D. Pfahl, "Reporting guidelines for controlled experiments in software engineering". in Guide to Advanced Empirical Software Engineering, F. Shull, J. Singer, D. Sjoberg, Eds. London: Springer, 2008, pp. 201–228.

[21] D. Cruzes, T. Dybå. Recommended Steps for Thematic Synthesis in Software Engineering. International Symposium on Empirical Software Engineering and Measurement, ESEM'11, September 2011, Banff, Canada, pp. 275-284.