UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UNDERSTANDING WEB SEARCH PATTERNS THROUGH EXPLORATORY SEARCH AS A KNOWLEDGE-INTENSIVE PROCESS

Marcelo Tibau de Vasconcellos Dias

Supervisors
Sean Wolfgand Matsui Siqueira
Bernardo Pereira Nunes

RIO DE JANEIRO, RJ – BRASIL
JANEIRO DE 2019

MARCELO TIBAU DE VASCONCELLOS DIAS

UNDERSTANDING WEB SEARCH PATTERNS THROUGH EXPLORATORY SEARCH AS A KNOWLEDGE-INTENSIVE PROCESS

Natureza do trabalho apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Estado do Rio de Janeiro, como pré-requisito para a obtenção do grau de Mestre em Informática.

Supervision:
Sean Wolfgand Matsui Siqueira
Bernardo Pereira Nunes

Rio de Janeiro

2019

UNDERSTANDING WEB SEARCH PATTERNS THROUGH EXPLORATORY SEARCH AS A KNOWLEDGE-INTENSIVE PROCESS

Marcelo Tibau de Vasconcellos Dias

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓSGRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Bernardo P.

Bernardo Pereira Nunes, D.Sc – UNIRIO / PUC-Rio

Sean W. M. Siqueira

Sean Wolfgand Matsui Siqueira, D.Sc – UNIRIO

Adriana Cesario Alvim

Adriana Cesario de Faria Alvim, D.Sc – UNIRIO

Luiz André Paes Leme

Luiz André Portes Paes Leme, D.Sc – UFF

RIO DE JANEIRO, RJ - BRASIL

JANEIRO DE 2019

Catalogação informatizada pelo(a) autor(a)

Dias, Marcelo Tibau de Vasconcellos
D541 Understanding Web Search Patterns Through
Exploratory Search as a Knowledge-Intensive
Process. / Marcelo Tibau de Vasconcellos Dias. --
Rio de Janeiro, 2019.
87

Orientador: Sean Wolfgand Matsui Siqueira.
Coorientador: Bernardo Pereira Nunes.
Dissertação (Mestrado) - Universidade Federal do
Estado do Rio de Janeiro, Programa de Pós-Graduação
em Informática, 2019.

1. Information Systems. 2. Exploratory Search.
3. Knowledge-intensive Process. 4. Query States. 5.
Query Reformulation. I. Siqueira, Sean Wolfgand
Matsui, orient. II. Nunes, Bernardo Pereira,
coorient. III. Título.

To my wife, Juliana.

ACKNOWLEGMENTS

I would first, and foremost, like to thank my advisors, Prof. Dr. Sean Wolfgand Matsui Siqueira of the Programa de Pós-Graduação em Informática (PPGI) at Universidade Federal do Estado do Rio de Janeiro (UNIRIO) and Prof. Dr. Bernardo Pereira Nunes of Programa de Pós-Graduação em Informática (PPGI) at Universidade Federal do Estado do Rio de Janeiro (UNIRIO) and Departamento de Informática at Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). The doors, e-mails, text messages, phone and video calls to Profs. Siqueira and Nunes were always open and available whenever I ran into a trouble spot or had a question about my research or writing. They consistently allowed this thesis to be my own work, but steered me in the right direction whenever I needed it. My gratitude for all the knowledge and experience shared and for showing me how to conduct a scientific research.

I would also like to offer my deep gratitude to Prof. Dr. Ivana Marenzi of L3S Research Center at Leibniz University of Hannover (Germany) and Prof. Dr. Maria Bortoluzzi of Udine University (Italy) for their unconditional support, sharing important data to this research as well as providing advises and paper reviews that deeply contributed to my formation as a researcher.

I am also thankful to the immensurable help of Prof. Dr. Fernanda Baião of Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), in which discipline this research originated. Her challenging constructively guidance helped me see the possibilities of this work.

I would like to extend my appreciation and gratitude to Ruben Francisco Manrique of Universidad de los Andes (Colombia) for all the aid provided to my research, specially its second cycle. I deep cherish the discussions we had and his valuable comments about this work.

I am also grateful to Philipp Kemkes of L3S Research Center at Leibniz University of Hannover (Germany) which was of paramount importance to the validation of the Exploratory Search KiP model, lending me his expertise and work that enable its applicability in less specific situations.

In the same way, I would also to thank my colleagues at Semantics and Learning Research Group and fellow students at PPGI UNIRIO for their unvaluable friendship and support throughout my Master's Degree. Andre Farzat, Bruno Crotman, César Augusto Rangel Bastos, Cleber Pinelli Teixeira, Crystiam Kelle Pereira e Silva, Davi Faisca Duarte, Jackson Luan Queiroz, Jerry Medeiros, João Paulo Baptista Voigtlaender, Luís Henrique Neves Villaça, Natália Joana Silva de Oliveira, Paulo Diogo Rodrigues Leão, Rafael Ferreira de Araujo, Reisla D'Almeida Rodrigues, Ronney Moreira de Castro, Rubia Silveira de Almeida, Solange Santolin and Thiago Baesso Procaci are some I would like to cite namely.

Finally yet importantly, I am forever indebted to my wife, Juliana, for being a true partner and friend in our joint path in this life. Your support and wisdom has been paramount to this thesis and to my improvement as a human being. To my parents, Marcio (*in memoriam*) and Euceli, for never being cheap on love, education and culture, providing me with books, courses, emotional support and expenses that helped made me who I am. In addition, to my extended family of brother, sister, in-laws, nephews, uncles,

aunts, cousins and friends that provided the necessary sense of belonging that makes one's life meaningful.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

DIAS, Marcelo Tibau de Vasconcellos. **Understanding Web Search Patterns Through Exploratory Search as a Knowledge-Intensive Process.** UNIRIO, 2019. 87 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

À medida em que os dados na Web crescem exponencialmente e abrangem a maior parte do conhecimento produzido pela humanidade, a busca por sistemas de informação inteligentes capazes de irem além das buscas por palavras-chave aumenta. Para entender melhor a intenção dos usuários, as ferramentas de busca da Web precisam transcender sua utilidade de classificadores de informações e aumentar seu discernimento semântico. Ao ajudar os indivíduos a recuperar informações relevantes através da formulação e modificação de suas consultas, os mecanismos de busca da Web ganham importância como ferramentas capazes de associar o aprendizado informal e o auto-aprendizado à aprendizagem formal. Neste novo contexto em que ferramentas de busca podem servir como meio para se obter conteúdo que suporte o aprendizado, tem-se a demanda de que sua atuação envolva uma capacidade de entendimento e processamento de informações mais complexas. Modelá-las para buscas que possuam um padrão mais exploratório, em que a informação disponível precise ser explorada em mais detalhes, é um caminho válido a se percorrer. Não sendo o escopo dos mecanismos de busca da Web atuais, a busca exploratória apresenta oportunidades para que se clarifique as melhores práticas associadas ao processo de tomada de decisão dos usuários em relação a informações adequadas e inadequadas e se utilize estes comportamentos para melhorar o entendimento do sistema de informação em relação ao propósito do usuário e na visualização dos padrões e do processo de aprendizado utilizado na própria busca. Nesta dissertação é apresentado o modelo KiP de Busca Exploratória, que ajuda a esclarecer as razões pelas quais um assunto é buscado e dá suporte à visualização de como as informações recuperadas são usadas para definir critérios de decisão sobre quais dados valem a pena serem salvos, fazer inferências e criar atalhos para a compreensão. Também é apresentada a Taxonomia ESKiP de Estados de Consulta, uma estrutura de classificação validada em um experimento em dois diferentes conjuntos de dados de *logs* de consulta e que auxilia na representação da estrutura e do comportamento da reformulação de consulta nos sistemas de busca. Representação esta que pode ser utilizada para configurações que organizem e forneçam resultados mais precisos.

Palavras-chave: Sistemas de Informação, Busca Exploratória, Processo-intensivo em Conhecimento, Estados de Consulta, Reformulação de Consulta.

ABSTRACT

As data grows exponentially and the Web encompasses most part of the knowledge Human Beings create the seeking for intelligent information systems capable of going beyond keyword searches increases. To better understands users' intent, Web search engines need to transcend its information sorter utility and acquire a more relevant ability concerning semantics' discernment. By aiding individuals to retrieve relevant information through formulation and modification of their queries, Web search engines gain importance as tools capable to associate informal and self-learning to formal learning. In a new context in which Web search engines can serve as a means to obtain content that supports learning, there is the demand that their performance involve a capacity for understanding and processing information that is more complex. Modeling them for searches with a more exploratory pattern, in which the available information needs to be explored in more detail, is a valid path to follow. Not being the scope of current Web search engines, exploratory search presents opportunities to clarify the best practices associated to users' decision-making process regarding suitable and not suitable information and to use these behaviors to improve information system's understanding about user's search purpose and to enable the visualization of Web search pattern and learning process used in the search itself. This thesis presents the Exploratory Search KiP model, which helps clarify the reasons why a subject is searched and supports the visualization on how the information retrieved is used to define decision criteria about which data is worth extracting, to draw inferences, and to create a shortcut to understanding. It also introduces the ESKiP Taxonomy of Query States; a classification framework validated in an experiment in two different query logs datasets and that helps to represent the structure and behavior of query reformulation in search systems. A representation that can be used to Web search engines' setups that organize and provide search results that are more accurate.

Keywords: Information Systems, Exploratory Search, Knowledge-intensive Process, Query States, Query Reformulation.

TABLE OF CONTENT

TABLE OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Contextualization

Exploratory search, the specialization of information seeking that combines activities of querying and browsing in order to obtain information [75] emerges as one of the fundamental skills in an environment that has the digital technology as one of the great driving forces [30, 31, 39, 62]. Its exploratory pattern derives from the fact that the user undergo complex cognitive tasks to explore the retrieved information and may lead to learning and acquisition/development of intellectual skills [75]. The contemporary digital technology's power[1] [31, 62] introduces different challenges to the tool in which exploratory searches are more commonly performed such as the Web search engines. To cite two of these challenges: a) the growing demand of solutions that allows information to be found in a resource-efficient way to make it available to the user in an appropriate context; and, b) the ability to identify users' specific needs or intentions to tailor search results to them.

In a scenario that presents no single right approach for a solution and in which problem definitions change as new information is gathered, the joint of exploratory search and the hugeness of data available added by increasing costs of education systems

[1] Manifested mainly in four aspects: big data, artificial intelligence & machine learning, cloud computing and Internet of Things.

worldwide [11, 29, 52] also insert the opportunity of using Web search engines as tools to support learning.

Learning from the search process, through the search activities performed, is a concept present in Search as Learning [71] and Information Seeking [36]. As part of their research agendas, various search tactics and patterns are explored in order to check how and what relationships between concepts arise. Those relationships may indicate a type of restructuration on user's knowledge frames promoted by the search and the interaction with the information retrieved. When the user analyzes the content, he/she can create a conceptual understanding about the topic searched. Clues that indicate a learning process in the making can be inferred as the user explores lists of retrieved documents through links, titles, abstracts, navigated texts, watched videos, among other possible data documents.

By modifying concepts and the relationship between them, the user might instigate a learning experience. Adding evaluative capacity and deepening his/her understanding of the subject [4], as a result. As commented by Vakkari [71], it is possible to recognize these new concepts added to the user's mental model by analyzing the search terms used. The number and specificity of the search terms tends to grow as the number of search sessions increases [72]. Again, as Vakkari [71] emphasizes, "growth of knowledge means growth in the number and specificity of concepts and their interrelations" and is shown by the way queries evolve during the search, for example, by the specialization of search terms.

Exploratory search is a core principle behind the use of search activities for learning purposes. Although a great number of Web searches could be considered exploratory, in the context of this study to be exploratory a search must transcend the intention to look for a simple data in the Web, such as the address of a particular establishment. It is also necessary to explore in more detail the information provided, i.e. the results returned by Web search engines or links available in portals. According to White and Roth [75], those searches are motivated by complex information problems and accompanied by misunderstandings about terminology and information structure. While users usually do not have enough previous information to help them to define a structured query, their own searching skills and knowledge of the context contribute to narrowing this gap [23]. This understanding leads to a paradigm review. If learning could be considered the main goal of an online search activity, then the quality of the results provided by Web search engines in terms of use effectiveness should prevail over keyword-document matching [76].

The quantity and relevance of retrieved items by Web search tools does not necessarily improve or optimize learning probabilities as search engines relevance criteria is based most on matching a document against a keyword. Moreover, users that do not have enough information to structure a proper query, as cited earlier in this Chapter, usually initiate exploratory searches. It is a situation known as Anomalous State of Knowledge or uncertainty [73], where [75]:

1) There is a lack of information to design a solution or to define the problem accurately;
2) There is a lack of an identifiable 'feasible solution' or an approach to deal with it;
3) Changes in the problem definition occur as new information is added.

Although lack of information is in itself a major challenge to any search, users' behavior and the tasks they choose to perform are of great importance to achieve a successful outcome as it facilitates a continuous flow of data gathering that can lead to learning. The research questions raised by these behaviors and tasks can be summarized in six topics:

- ✓ How the initial search terms are chosen?
- ✓ How the search terms are evaluated?
- ✓ How and why search terms are modified during a search session?
- ✓ How the results are checked?
- ✓ What is the Decision Process involved?
- ✓ What features determine data suitability?

In this sense, user's previous search experiences and background are pivotal to understand search intentions and information acquisition patterns. Under these circumstances, exploratory searches can be considered a Knowledge-intensive Process (KiP).

1.1.1. Knowledge-intensive Process (KiP)

KiPs are related to Business Process Management (BPM), Process Management Systems (PMS) and Knowledge Management (KM) [12]. BPM is an area of research used as an ongoing basis for business process improvement as it provides techniques and software to design, enact, control, and analyze business processes [12] in order to organize their activities and understand their interrelationships [74]. A process management are often based on "the assumption that processes are characterized by repeated tasks, which are performed on the basis of a process model prescribing the execution flow in its entireness" [37].

To automate the prescribed execution flow of a business process, the usage of technological tools that include human-focused tasks along with machine processing applications is necessary to allow an organization to manage its work flexibly. Those tools are known as PMS and model activities along three perspectives [28]:

1) A control-flow perspective, which describes a structure of a process in terms of tasks and the relationships between them;
2) A data perspective with the data elements consumed, produced and exchanged during the process execution;
3) A resource perspective that deals with the process' operational and organizational contexts in terms of resources.

KM is a multidisciplinary approach focused on the management of the way knowledge and information are created, shared and used throughout an organization [19]. Its main framework categorizes two distinguished dimensions of knowledge: tacit and explicit [44]. Tacit knowledge represents the internalized knowledge that an individual possess – and may or may not be consciously aware of – but uses to accomplish particular tasks. Explicit knowledge, on the other hand, represents knowledge that an individual holds consciously in a form that can easily be communicate to others. KM goal is to convert internalized tacit knowledge into explicit knowledge and make it available to transfer within a group.

Knowledge can be defined as a combination of experience, context, interpretation and reflection and, while considered as a process, it involves more human participation than information [9]. The knowledge-intensity is recognizable by the diversity and uncertainty of process input and output [10]. Di Ciccio, Marrella and Russo emphasize that this definition suggests that process-related knowledge is strongly human-centered [12] and deeply rooted on creation, co-creation, sharing, transferring and application of knowledge in the context of the processes that people participate.

Knowledge-Intensive Process Ontology (KIPO) enables to understand and represent a knowledge-intensive process rather accurately [17]. The relevance of the ontology in this study lies in exploring elements from tacit knowledge [17] and using them to represent KiP's features. KIPO fulfills this role well because it already has a specific graphical notation to represent tacit knowledge, the Knowledge-Intensive Process Notation (KIPN). This notation is presented and explained thoroughly in [42, 43], but here we summarize the main features related to a set of diagrams used to represent search and user perspectives within a KiP.

The first of these diagrams is the KiP Diagram, which indicates constraints to the flow as well as represents circumstantial events, innovations and decisions. Another diagram is the Socialization Diagram, which has the role of showing how knowledge acquisition and sharing take place within activities. There is also the Decision Diagram, which has the aim to match detailed decision-making processes with their respective results. Finally, there are three last sets of diagrams:

a) Agent Diagram that maps the agents' experience and expertise and illustrates their skills;
b) Intention Diagram that represents desires, feelings and beliefs that motivate an agent to engage in activities, decisions or socializations;
c) Business-Rules Diagram that represents documented business rules that limit

a decision during a Knowledge-intensive Process, e.g., legislation, contracts, and internal regulations.

Although informal regulations/conventions that are part of the professional knowledge of users are very relevant in the context of this work, we did not used the Business-Rules Diagram *per se*, but considered this type of tacit knowledge as part of users' Decision Diagram instead.

1.1.2. Query States

Exploratory search is an activity in which uncertainty plays a fundamental role and in which a berrypicking pattern [2], with bit-by-bit information gathering, is necessary to create systematic awareness about a situation or fact. A search quest can lead to many paths but not necessarily, those connected to learning. Exploratory search as a KiP involves visualizing its tasks as part of a chain of actions that, consistently performed, helps to clarify reasons behind search tasks, meanings attributed to the information retrieved and how it is used to promote user's awareness about the subject. This is closely tied to the definition of exploratory search, shared by authors such as Vakkari [72, 73], White and Roth [75] and Wildemuth and Freund [76]. In order to develop more assertive Web search engines that can be used as learning tools, it is necessary to view knowledge as an integral part of the process. One possible manner to implement this goal is to use the formulation and development of search queries as a parameter of the search evolvement.

The use of Web search engines to conduct search tasks on the internet is a type of communication interaction known as P2M (Person-To-Machine), where humans start, change or end an operation on a machine. Query states are used to model the sequence of

P2M interaction during a Web search [7, 25]. The state of a query is based on the terms used in consecutive queries and is determined by query reformulation. As defined by Jansen, Booth and Spink [25], query reformulation is the process of altering a given query to improve search or retrieval performance.

Also known as query expansion and query modification, query reformulation is a trend subject of research for the past few decades, since Gauch and Smith has shown in 1993 [18] that effective query reformulation can improve the outcome of users' searches. According to Rha, Shi and Belkin [16], the way users choose to transition queries has been a major focus of research, particularly regarding specific types and patterns of query reformulations such as add and remove words, words substitution and spelling correction.

1.2. Methodological Approach

The exponential growth of data on the Web encompass the challenge of having intelligent search systems that transcend their approach of sorting and organizing information by indexing keywords. Web search engines are configured to retrieve information based on keyword-document match, which does not add much to searches influenced by information acquisition patterns like exploratory searches.

If we could understand how users form their information gathering behavior, then this knowledge could be used to build information systems capable not only to retrieve relevant information but also to understand a search pattern and adjust the results accordingly. As consequence, the Web search engine could retrieve information that is more relevant and provide results that are more accurate based on the users' intent.

In this Master Thesis, the Exploratory Search KiP model is presented. It is an exploratory search model that helps understand the information-gathering behavior embedded in an exploratory search and visualize the knowledge process by mapping the search's Intention and Decision Diagrams. The model was designed and developed using Design Science Research principles, which will be further explained in Chapter 2. This artifact was devised and evaluated in a collaboration among Italian, German and Brazilian Universities. The first research cycle[2] (Figure 1) was carried out applying the proposed model in a case study scenario with six specialist-teachers and validating with an in-depth analysis of the results produced by interviews based on the think-aloud protocol[3] analysis to validate it. The model was re-validated to check its applicability in less specific situations than those provided by the specialist-teachers, by search log analysis from an online professional community collected during a workshop, in which 22 users performed 1,249 search sessions.

[2] The creation and evaluation of artifacts forms an important part in the Design Science Research process, which develops them in closely related activities during research cycles.

[3] The protocol involves that participants describe aloud what they are thinking while performing a set of specified tasks.

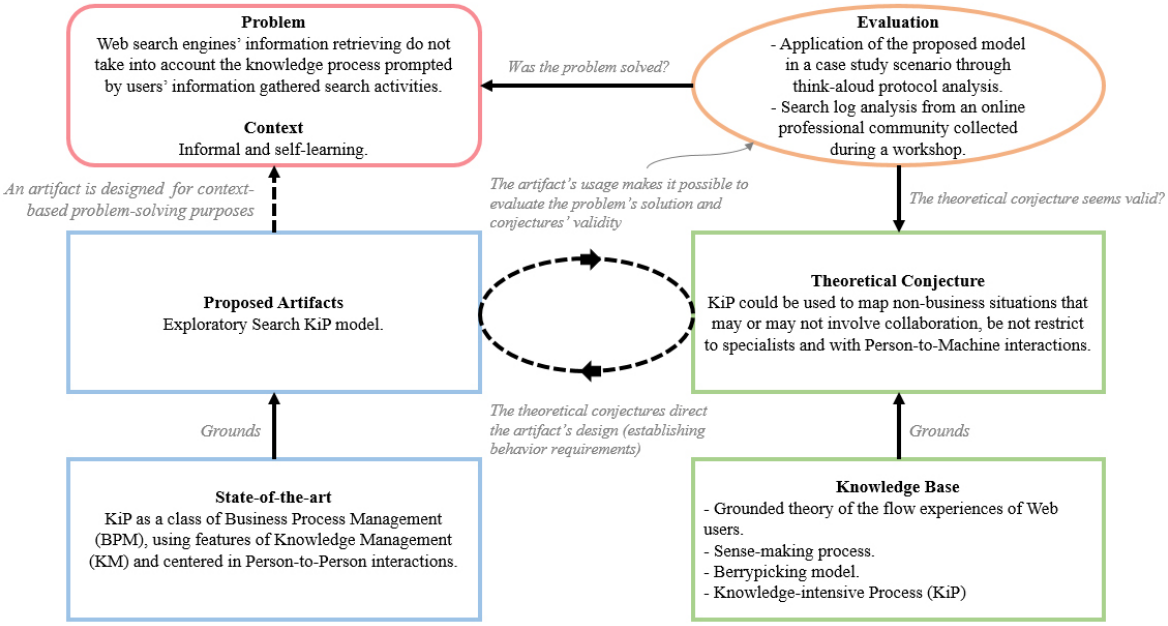
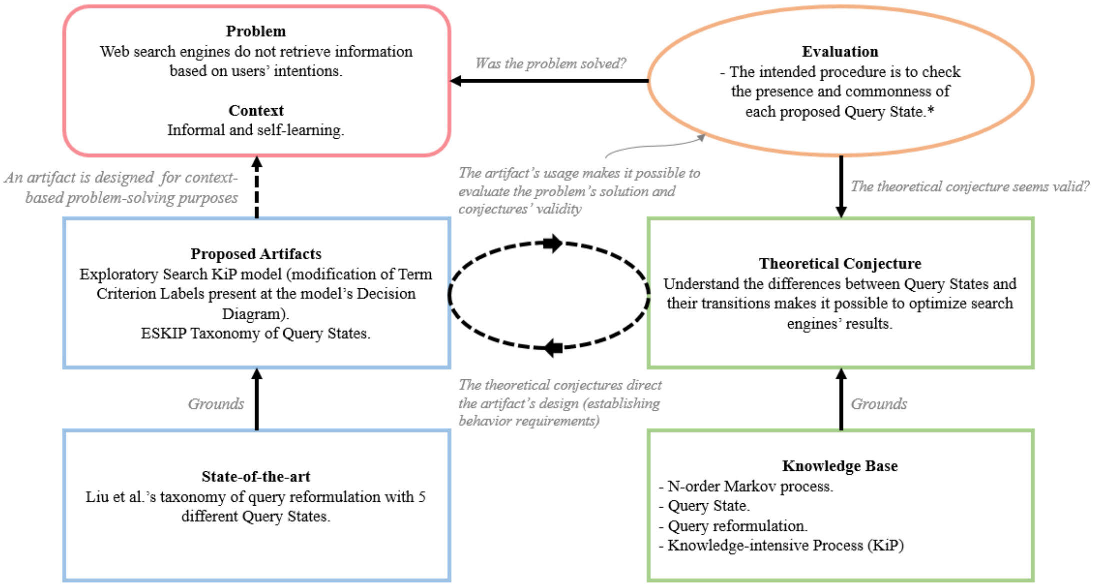

Problem
Web search engines' information retrieving do not take into account the knowledge process prompted by users' information gathered search activities.
Context
Informal and self-learning.
Was the problem solved?
Evaluation
- Application of the proposed model in a case study scenario through think-aloud protocol analysis.
- Search log analysis from an online professional community collected during a workshop.
An artifact is designed for context-based problem-solving purposes
The artifact's usage makes it possible to evaluate the problem's solution and conjectures' validity
The theoretical conjecture seems valid?
Proposed Artifacts
Exploratory Search KiP model.
Theoretical Conjecture
KiP could be used to map non-business situations that may or may not involve collaboration, be not restrict to specialists and with Person-to-Machine interactions.
Grounds
The theoretical conjectures direct the artifact's design (establishing behavior requirements)
Grounds
State-of-the-art
KiP as a class of Business Process Management (BPM), using features of Knowledge Management (KM) and centered in Person-to-Person interactions.
Knowledge Base
- Grounded theory of the flow experiences of Web users.
- Sense-making process.
- Berrypicking model.
- Knowledge-intensive Process (KiP)

Figure 1: DSR elements' mapping on the first Research Cycle.

The first research cycle had its problem stated as "Web search engines' information retrieving do not take into account the knowledge process prompted by users' information gathered search activities". The statement relates to the necessity of use effectiveness prevailing over keyword-document matching as the main Web search engines' results parameter, a comment made at section 1.1. As shown in Figure 1, KiP was applied to a non-business situation, exploratory search, in which some important aspects such as being heavily dependent on knowledge workers, taking place in a collaborative multi-user environment and P2P type of interaction had to be adapted.

A taxonomy of Query States is also introduced. This work equally relied on Design Science Research principles and derived from the knowledge obtained by the use of Exploratory Search KiP model, especially the insights provided by the decision-process mapping. A collaboration among Colombian and Brazilian Universities, the ESKiP Taxonomy of Query States, as it was named, is a classification framework that helps to represent the structure and behavior of query reformulation in search systems.

In addition to the taxonomy, this second research cycle (Figure 2) introduced a modification to the Exploratory Search KiP model. Its problem statement alludes to a challenge to Web search engines cited likewise in the previous section, the ability to tailor its results based on users' specific purposes. The taxonomy was validated in an experiment in two different query logs datasets in which the intended procedure was to check the presence and commonness of each proposed Query State to verify the taxonomy's representational performance.


Problem
Web search engines do not retrieve information based on users' intentions.
Context
Informal and self-learning.
Was the problem solved?
Evaluation
- The intended procedure is to check the presence and commonness of each proposed Query State.*
An artifact is designed for context-based problem-solving purposes
The artifact's usage makes it possible to evaluate the problem's solution and conjectures' validity
The theoretical conjecture seems valid?
Proposed Artifacts
Exploratory Search KiP model (modification of Term Criterion Labels present at the model's Decision Diagram).
ESKIP Taxonomy of Query States.
Theoretical Conjecture
Understand the differences between Query States and their transitions makes it possible to optimize search engines' results.
Grounds
The theoretical conjectures direct the artifact's design (establishing behavior requirements)
Grounds
State-of-the-art
Liu et al.'s taxonomy of query reformulation with 5 different Query States.
Knowledge Base
- N-order Markov process.
- Query State.
- Query reformulation.
- Knowledge-intensive Process (KiP)

*determine its representational adequacy by validating the epistemic requirement (the representation must provide the ability to express the facts it wish to express) of Knowledge Representation Schemes evaluation.

Figure 2: DSR elements' mapping on the second Research Cycle.

To determine its representational adequacy, the first step of the evaluation cycle attempted to validate at the first dataset the epistemic requirement of Knowledge Representation Schemes evaluation. Bingi, Khazanchi and Yadav [3] defines the requirement, asserting that the representation must provide the ability to express the facts it wish to express. To observe the Query States' presence and commonness, each query reformulation type was counted and its overall frequency verified. The second step of evaluation attempted to check the proposed Query States' consistency by verifying its appearance in a different dataset of queries. Again, the query reformulation types had their presence asserted by examining their frequency.

1.3. Contributions

The main contributions of this Master Thesis are the two artifacts introduced in section 1.2: (i) the Exploratory Search KiP model; and, (ii) the ESKiP Taxonomy of Query States.

The Exploratory Search KiP model provided an update to the use of Knowledge-intensive Process with its adaptation to a non-business situation. It has the ability to present not only the information-gathering behavior embedded in an exploratory search, but also the knowledge process prompted by the information gathered and shown in its Diagrams.

The ESKiP Taxonomy of Query States produced an update to the state of the art of Query State transitions and was applied to a classifier that automatic identifies query reformulations. In addition, it provides as technical contribution two queries datasets with its Query States identified and labeled.

As part of Searching as Learning's research agenda, the two proposed artifacts are relevant contributions to the development of searching systems that could support critical and creative learning through search behavior. Rieh *et al.* [59] exemplifies some of those behaviors such as the abilities of evaluating the usefulness of information critically, differentiating resources, monitoring results, tracking information, prioritizing actions and applying sense-making. The contribution resides on the artifacts' capacity to represent both the knowledge process derived from the user's information-gathering behavior and the relationships between concepts based on the query terms used.

1.4. Thesis Outline

The remainder of this thesis is structured as follow: Chapter 2 explores the first research cycle, explaining how the DSR method was applied to the Exploratory Search KiP model's development and discusses the model's generalization by sharing insights from its application to log analysis. Chapter 3 presents the second research cycle, explaining ESKiP Taxonomy of Query States, its application to an automatic classifier and the classifier's evaluation by using string similarity. Finally, Chapter 4 concludes the thesis and deliberates about the research's future developments.

2. MODELING EXPLORATORY SEARCH AS A KNOWNLEDGE-INTENSIVE PROCESS

2.1. Introduction

To consider a Web search engine as a tool able to transcend its information sorter utility and acquire a more relevant role concerning the learning process of an individual, it is important not to rely only on models that focus on interactive information retrieval, as discussed by Wildemuth and Freund [76]. In this respect, it is important to bear in mind that the use of the term “learning”, considering Searching as Learning’s concepts and our approach of them in this research, does not relate to the learning theories that might stimulate the acquisition of knowledge or the development of higher levels of intellectual capacity. It relates to learning’s role as a search intention as well as a consequence of a Knowledge-intensive Process.

The growth in the number and specificity of concepts and their interrelations provided by the addition of information in each stage of a search[4], as cited by both Vakkari [71] and Bates [2], might play a role in the increase of user’s capacity to apply, synthesize and evaluate the search terms used. It is important to state that it is not a focus of this study at the present stage to propose ways to account the addition of information provoked by a search as a measure of learning. The conceptual framework that could account for

[4] Usually observed by the specialization of subject domain.

the processes of learning that occurred during searches depends on a better understanding about human behavior and cognition efforts while using Web search engines as learning tools and goes beyond the determined attempt of this research. What is expected to be attained is to successful identify transitions in query reformulations, in order to use this knowledge, in a further step, to build information systems capable to understand Web search patterns and adjust the results accordingly, contributing to the user's search and learning processes.

As computers become ubiquitous and the internet is the locus of convergence for mass media, searching the Web becomes more than a habit for a wide range of people. In this sense, defining exploratory search is no easy task, since the majority of searches are, in some way, exploratory. The definition laid on the first chapter provided the degree of distinctiveness necessary to outline our approach of viewing exploratory search as a KiP and to use this understanding to adapt a business process model, as KiP is, to a non-business situation, as exploratory search is.

2.2. Related Work

The study of information seeking and exploratory searches and processes has been of interest in a number of areas such as human-computer interaction, information retrieval, information science and psychology. Such areas have different views of exploratory search and has helped in its evolution and understanding. For instance, understanding and modeling emotions like curiosity and uncertainty along with exploratory behavior and browsing patterns help understand cognitive and reactive expressions of human beings.

Pace [51] proposes a psychological theory to exploratory search, the grounded theory of the flow experiences of Web users, adapting a categorization of different types of curiosity to information seeking which is also relevant to exploratory search: it is the difference between diversive curiosity and specific curiosity. The first related to a general seeking of stimulation or novelty, analogous to television channel surfing behavior, and leads to ill-defined goals and exploratory browsing when under an exploratory search process. The latter can be summarize as a desire for a particular piece of information or subject and leads to a well-defined goal and direct searching. The different types of curiosity, linked to the balance between the challenges of the activity and the skills required to meet those challenges, define the search success.

Curiosity and the tendency to engage in information seeking behavior might depend on individual differences and account for what Pace [51] reported in his study as two general types of navigation attitudes. One is a directed searching mode in which the user is motivated to find a particular piece of information and the other, an exploratory browsing mode with a more empirical nature. In the approach of exploratory search outlined for the present research, it was noticed two possibilities of search goals defined beforehand by users: learning and investigation. Learning, as an exploratory search process goal, occurs if the user’s intention is to expand his/her level of knowledge on a given topic. Investigation occurs if the intention is to acquire a more superficial level of knowledge. It is important to emphasize that the cited “more superficial level of knowledge” does not concern a trivial search, such as look for an establishment´s address as cited in Chapter one. It concerns searches that do have an exploratory pattern and that do demand complex cognitive tasks to explore the retrieved information, though not as meticulously as it would be if the intention were learning.

Similarly to Pace [51], the present research also looked at the different types of curiosity which are important to define search goals (learning or investigation) and search behavior. However, unlike Pace's theory, the present work takes into consideration the lack of information of a user at the search problem definition or solution, which exists at the beginning of a search. Pace [51] argued that Web users engaged in directed searching have a clearly defined goal, and those engaged in exploratory browsing an ill-defined one. In the view considered in the current study, it is not a matter of better or poorly defined goals but a gap between the search intention and the search savvy necessary to define the proper shortcut towards understanding. The important point is not the presence of a goal in each case, as Pace [51] emphasizes, but rather a mixture of previous searching expertise and skills to define structured queries and searching strategies to meet those goals.

Variants of curiosity and their roles in search are also present in sense-making theory [26, 33, 56, 61]. This particular theory studies the creation of situational awareness and understanding in situations of uncertainty, high complexity or involving stochastic processes in order to take decisions. The exploratory search process also uses a degree of sense-making in its observation and cognitive task analysis to define decision criteria. Pirolli and Card [56] propose a model based on the theory that could be analogous to the one presenting in this thesis. Both models drawn toward the role of knowledge in the process as decisive.

Pirolli and Card's sense-making process [56] consists in tasks of information gathering, re-representation of the information in a schema that aids analysis, the development of insight through the manipulation of this representation, and the insight's application to create some knowledge product or to direct action. This process is summarized in a formula: Information -> Schema -> Insight -> Product and is based on Russell *et al.*'s "learning loop complex" [61]. The Learning Loop Complex theory of sense-making is composed of three loops, the first one is the “generation loop” which is a search for a good representation. It is followed by an attempt to encode information in the representation named “representational shift loop”. When the attempt at encoding information in the representation identifies items that do not fit, it generates a residue and originates an attempt to adjust the representation to provide a better coverage, called the “data coverage loop”.

This information processing has a two-way driven force: bottom-up processes, from data to theory or top-down, from theory to data. The bottom-up approach is characterized by activities as search and filter, read and extract, schematize – where the information is re-represented in some schematic way, build case – where a theory or case is built by additional evidence to support or disconfirm it, and tell story – which is the presentation or publication of the theory/case. Note that the activities “search and filter” as well as “read and extract” are also considered exploratory search activities. The top-down approach is composed of re-evaluation activities to adjust and search for support, evidence and relations.

Despite the fact that the model proposed in the present thesis and sense-making could be comparable in certain respects, exploratory search as a KiP is not limited by sense-making waterfall approach of progression data to information, knowledge and understanding. As a Knowledge-intensive Process this pattern is considered synchronous, meaning that the referred progression could occur at the same time or in parallel. Thus, exploratory search takes into account user's previous search experience and search intention to define decision criteria to data suitability as well as a tool to draw inferences and create a shortcut to understanding.

Another approach related to this work is the berrypicking model [2]. The analogy to picking berries in a forest, in which they must be picked singly, is fitting to exploratory search process since the approach views the user moving through an information space in a similar way, gathering fragments of data and seeking cues to decide over information suitability and navigation decisions.

Berrypicking model [2] works within a type of search called evolving search that usually begins by users with just one feature of a broader topic or just one relevant reference. The user then proceeds to move through a variety of sources, adding new piece of information while using it to define directions and search strategies to follow. At each stage, the user is not just modifying the search terms used in order to get a better match for a query, but the query itself through the search terms used. Moreover, with each different conception of the query, the user identifies useful information and references in a way that the query is not satisfied by a single final retrieved set of information, but by a series of selections of individual references and bits of information at each stage of the ever-changing search in what Bates [2] called berrypicking pattern.

The evolving search is in essence an exploratory search, with its characteristic exploratory pattern of complex cognitive tasks used to explore the retrieved information. The only difference is the appearance provided by the analogy of picking huckleberries or blueberries that have to be scattered on the bushes since they do not come in bunches. The information's berrypicking is not evolving by itself, as Bates acknowledges, as one could do berrypicking without the changing of the search's need.

Although Bates [2] states that "new information encountered gives the searcher new ideas and directions to follow and, consequently, a new conception of the query", it also emphasizes the variety of search techniques used and the nature of the search process – berrypicking pattern – rather than the search process itself. In this sense, exploratory search as a KiP updates the berrypicking approach and its iterative behavior, especially in its query variation based on applying thoughts on documents and information retrieved. Exploratory search as a KiP considers the transformation of desire into intention to form a mental image that changes, adapts and evolves throughout the process, as well as its influence on search term selection and refinement and on decision criteria regarding link examination and data object extraction.

The study of Business Processes, their managements and frameworks, is a key element to comprehend how activities intertwine and how their relationships are used to improve understanding about human, organizational, documental and other sources of information interaction within them [74]. Knowledge-intensive Process is a class of Business Process Management (BPM) developed to assist the mapping of processes' knowledge dimension and to consider the role of human-centered knowledge in shaping process' activities [12].

KiPs are characteristically less rigid in structure and usually involve autonomous user decisions and unpredictable events. Its importance in the BPM domain has emerged due to a prominent role of knowledge workers – a type of worker whose main capital is knowledge such as engineers, scientists, lawyers, academics, among other white-collar workers – play in modern organizations [12, 13, 58]. As cited by Di Ciccio, Marrella and Russo in [12] and [13], KiPs are "often related to the need of considering and understanding the knowledge dimension in business processes". As consequence of the shift from an industrial society to a knowledge-based society, the role of knowledge in a process goes beyond its traditional approach that intends to manage processes and process-related knowledge separately to an approach that considers it an integral part or the process itself.

Vaculin *et al.* [69] provide a definition accepted by other researchers in BPM field such as Di Ciccio, Marrella and Russo [12], concerning the main characteristics of this particular class: KiPs are “processes whose conduct and execution are heavily dependent on knowledge workers performing various interconnected knowledge intensive decision making tasks. KiPs are genuinely knowledge, information and data centric and require substantial flexibility at design- and run-time”.

The research presented in this thesis do agree with most characteristics derived from this definition such as knowledge-driven, unpredictability, emergence, goal-oriented, event-driven, non-repeatable, constraint and rule-driven. However, the collaboration-oriented characteristic is regarded as too restricted for the scope of a Knowledge-intensive Process applied to the current study. Di Ciccio, Marrella and Russo [12] defined this particular characteristic as “process creation, management and execution” occurring in a “collaborative multi-user environment, where human-centered and process-related knowledge is co-created, shared and transferred by and among process participants with different roles”. Creation, co-creation, sharing and transferring of knowledge are features drawn from Knowledge Management (KM) concepts [38] and are centered in P2P (Person-to-Person) interaction, especially involving people with specialist domain.

While applying KiP to exploratory search, one more role not considered by its KM inheritance was perceived: acquisition of knowledge. By the inclusion of this new role in Knowledge-intensive Process, it is possible to register that exploratory search may or may not involve collaboration among people. Certainly is not restricted to specialists as well, since it introduces as process' agent the non-specialist or lay person, which is the case of a user that starts a search without previous knowledge about the subject. In this respect, the new view updates KiP's characteristics by considering a non-collaborative aspect related to learning and disputing knowledge workers' ascendancy.

Another update to KiP provided by this new approach concerns the type of interaction itself. In addition to P2P, it can be P2M (Person-to-Machine), which is the case of using Web search engines to perform an exploratory search. A Web search engine acts as an intermediary between the user and the Web, sorting and retrieving information from internet documents based on queries and the terms used. While P2P interaction does not necessary require technology to come to fruition, P2M interaction thrives in it. This is precisely what makes this issue particularly important to consider while identifying, systematizing and implementing the knowledge-intensive approach in real-world processes that rely more and more on technology usage.

2.3. The Exploratory Search KiP Model

In order to visualize the Knowledge-intensive activities and the user's characteristics and behaviors, to model the actions using the KIPN notation diagrams and to propose an initial model encompassing exploratory search features, a series of searches were conducted based on tasks originally presented by Kules and Shneiderman [27], which were designed to map navigation and search patterns. The result of this earlier work was the definition of four exploratory search activities: (1) Search Term Selection, (2) Query Formulation, (3) Results Check, and (4) Information Extraction.

To explain the model's characteristics, it was opted to present it as a modeled search and describe the conceptual basis of each exploratory search activity separated by each KIPN diagram used. For visualization purposes, it was also generated a digital object identifier (DOI)[5] that shows the set of diagrams presented in each Figure. The diagrams include the tacit elements linked to the process. The description of each element used are described as follows:

KiP Diagram

- Knowledge-intensive Activities: meaning the Exploratory Search KiP activities involved in the process' P2M interaction;
- Fact: meaning some occurrence in the KiP scenario;
- Process goal: meaning an expected objective for executing the process.

Intention Diagram

- Desire: meaning something that the user desires;
- Intention: meaning something that the user intends to achieve when performing an activity;
- Mental image: meaning an interpretation and mental organization of information that creates knowledge.

Decision Diagram

- Criterion: meaning established criteria for analyzing alternatives' advantages and disadvantages;
- Evidence: meaning a proof or a sign that something exists;

[5] DOI number 10.13140/RG.2.2.20049.63840

- Restriction: meaning a law, rule, or any known circumstance that may restrict the decision-making;
- Question: meaning an issue considered by the user when making a decision;
- Decision: meaning the result of a decision-making process;
- Chosen alternative: meaning the selected alternative to address issues of a decision.

2.3.1. KiP Diagram

The model's overview (Figure 3)[6] shows the mentioned four exploratory search activities: Search Term Selection, Query Formulation, Results Check, and Information Extraction. The KiP diagram indicates the flow of activities and signals the possibility of synchronicity between them, an attribute of Exploratory Search KiP model. An example of this simultaneous occurrence of events is the search actions themselves, often performed not by terms and queries but by browsing through clicks and links or by selecting categories while searching on platforms such as online communities and portals. This suggests the possibility of parallelism between activities, e.g. refining the search just by looking at the results.

[6] It is worthwhile to comment that although the Figures exemplify the Exploratory Search KiP model, they are instances used for illustrative purposes rather than explanatory.

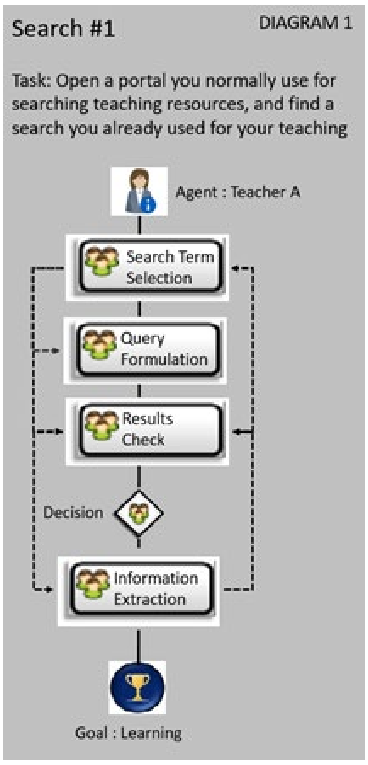


Figure 3: Overall Exploratory Search Knowledge-Intensive Process.

2.3.2. Intention Diagram

The Intention Diagram was integrated in each activity to help to visualize the flow of actions that lead to users' mental image's formation and refinement about the subject sought and its impact on the decision process. The choice to integrate it was encouraged by the necessity of better observe the desire – which motivated the search – transforming into intention, the action taken to put the search into practice and how this desire-intention interaction affected the domain's perception, characterized at Exploratory Search KiP model as mental image.

The Search Term Selection activity is a purpose unifier activity. Figure 4, encompasses what Ellis *et al.* [14] defined as Starting, information search forming activities and Chaining, initial source checking activities. Starting and Chaining activities are applied in order to form an initial desire – learning, if the intention is to expand the level of knowledge about a given topic or investigation, if the goal is to acquire some level of information. The symbols shown in the diagram represent a search desire, an intention – in the activity the search term selected – and the formed mental image from the initial checking of the information retrieved.

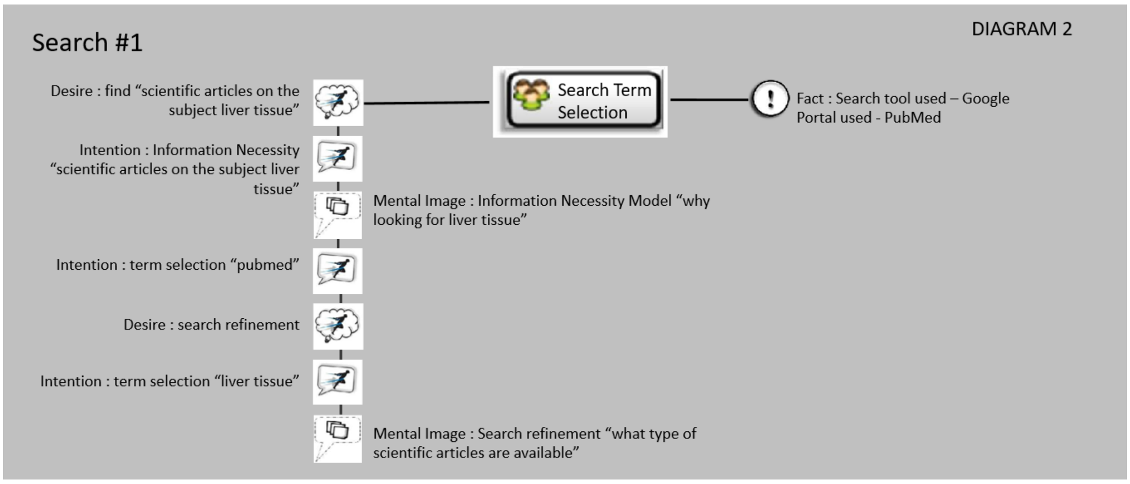

Search #1
DIAGRAM 2
Desire : find "scientific articles on the subject liver tissue"
Search Term Selection
Fact : Search tool used – Google Portal used - PubMed
Intention : Information Necessity "scientific articles on the subject liver tissue"
Mental Image : Information Necessity Model "why looking for liver tissue"
Intention : term selection "pubmed"
Desire : search refinement
Intention : term selection "liver tissue"
Mental Image : Search refinement "what type of scientific articles are available"

Figure 4: Search Term Selection activity (Diagram 2).

As expressed by Vakkari [73], an exploratory search usually starts with ill-structured problems as it begins with a lack of information necessary to develop a solution or properly define the problem. As previous seen, it also presents no single right approach for a solution and has problem definitions that change as new information is gathered. Query Formulation activity helps to structure the search and presents an initial doorway from the desire that prompted the search to a better-defined search intention. The activity involves search strings formulation and combines all the texts, numbers and symbols entered by a user into a search engine. When users use structured strings instead of keywords, they are able to direct search tools to return results that are also more structured, which increases the chance for the retrieved information to be more assertive regarding the search intention, thus enhancing suitability. Figure 5 attempts to capture how the formed mental image helps users in choosing the most appropriate filters to perform the search and in formulating a query that provides more useful results.

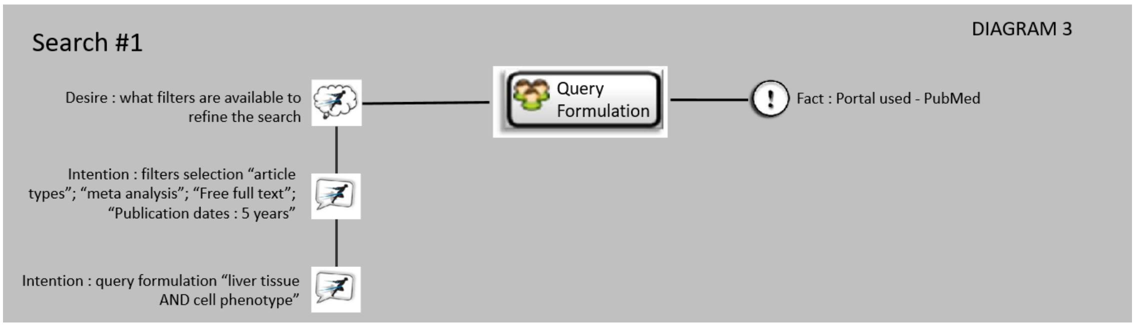

Search #1
DIAGRAM 3
Desire : what filters are available to refine the search
Query Formulation
Fact : Portal used - PubMed
Intention : filters selection "article types"; "meta analysis"; "Free full text"; "Publication dates : 5 years"
Intention : query formulation "liver tissue AND cell phenotype"

Figure 5: Query Formulation activity (Diagram 3).

The selection of filters characterizes a refinement through variables definition. Which is possible when users gain a better understanding about the subject and attempt to define initial criteria that could be used to start to assess the relevancy of the retrieved information.

Results Check activity, shown in Figure 6, returns to Ellis *et al.* [14] to perform:

- Browsing, semi-directed search through potential search locations;
- Differentiating, filter and select sources by quality and relevance;
- Monitoring, continued reviewing of sources identified as promising, as shown by the several links examined.

In this activity, users seek to formalize the initial set of criteria. The formalization helps them decide over the quality of the information retrieved, in this particular example, 'known sources and academic relevance'. Focused search, as described by White and Roth in [75], also applies in this activity once the examined user retrieved documents by clicking at certain links to decide over relevant information that meets the search intention.

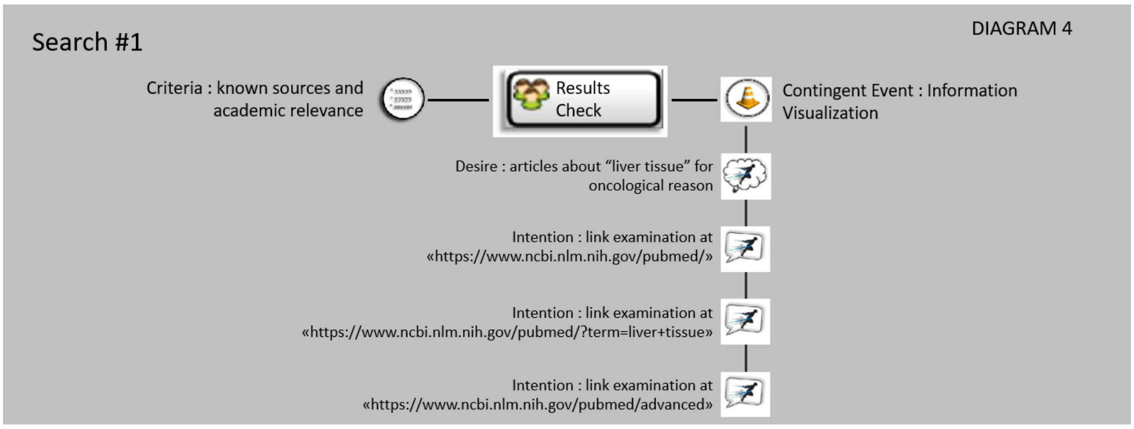

Search #1
DIAGRAM 4
Criteria : known sources and academic relevance
Results Check
Contingent Event : Information Visualization
Desire : articles about "liver tissue" for oncological reason
Intention : link examination at «https://www.ncbi.nlm.nih.gov/pubmed/»
Intention : link examination at «https://www.ncbi.nlm.nih.gov/pubmed/?term=liver+tissue»
Intention : link examination at «https://www.ncbi.nlm.nih.gov/pubmed/advanced»

Figure 6: Results Check activity (Diagram 4).

The symbols shown in Figure 6 represent defined criteria for the evaluation of the results' quality and a contingent event characteristic of the Results Check activity: information visualization. Results Check induces the user to acquire a clear sense of the information needed and the trails to follow to obtain it.

Information Extraction, the last activity presented for the Exploratory Search KiP model, is based on the work of both Ellis *et al.* [14] and Marchionini [36], undertaking actions to extract the data considered suitable using a reflect/iterate/stop process, as shown by the data object symbol in Figure 7. To a certain extent, the presented Information Extraction activity has a bearing on knowledge appropriation. Appropriation means, in this context, the process of transferring knowledge from sources to users. It modifies the user's cognitive processes of sense-making, which gives meaning to experience, and of decision-making, which results in selecting a course of action among several alternatives. As a result, one or more data objects defined by relevance assigned to viewed documents are created.

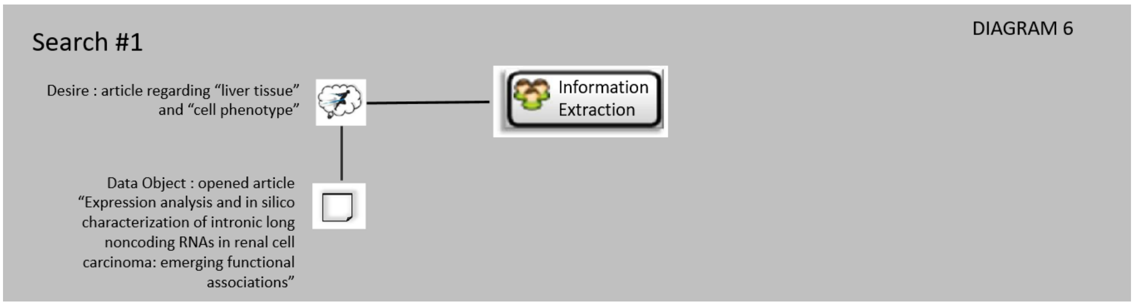

Search #1
DIAGRAM 6
Desire : article regarding "liver tissue" and "cell phenotype"
Information Extraction
Data Object : opened article "Expression analysis and in silico characterization of intronic long noncoding RNAs in renal cell carcinoma: emerging functional associations"

Figure 7: Information Extraction activity (Diagram 6).

2.3.3. Decision Diagram

To visualize users' choices based on the perceived criteria, the decision process used by them to determine the resources' appropriateness was carefully examined. The Decision Diagram shown in Figure 8 played a decisive role to represent how the search context and user expertise work together to define suitability, prompting the definition of two types of decision criteria to be detailed in Subsection 2.4.2.

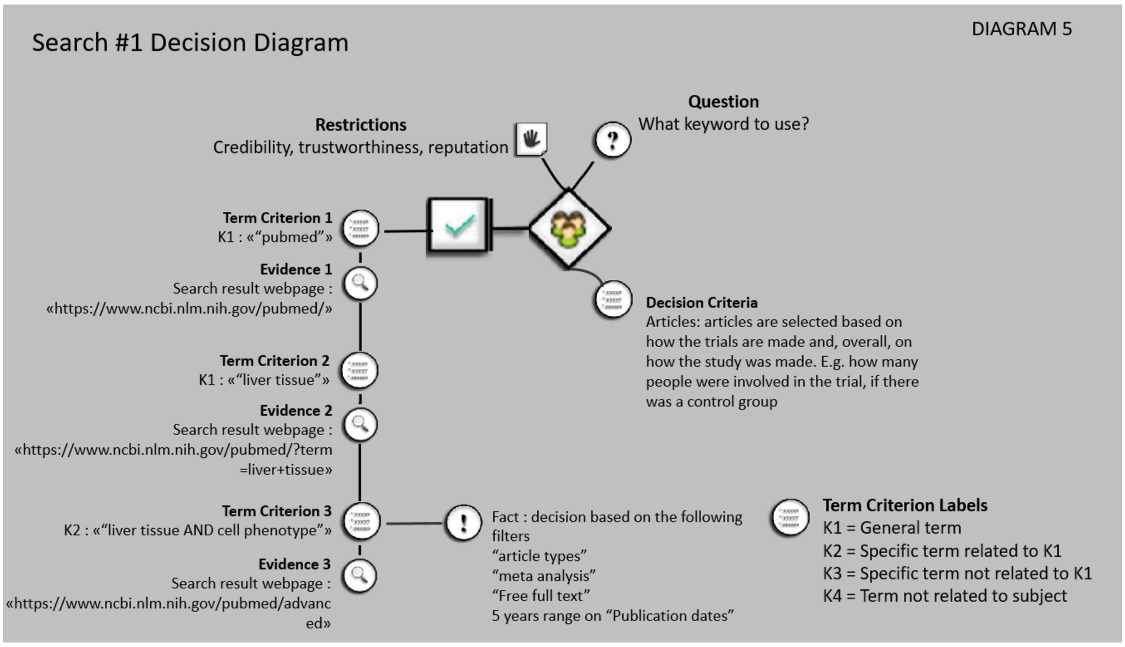

Search #1 Decision Diagram
DIAGRAM 5
Restrictions
Credibility, trustworthiness, reputation
Question
What keyword to use?
Term Criterion 1
K1 : «"pubmed"»
Evidence 1
Search result webpage : «https://www.ncbi.nlm.nih.gov/pubmed/»
Term Criterion 2
K1 : «"liver tissue"»
Evidence 2
Search result webpage : «https://www.ncbi.nlm.nih.gov/pubmed/?term =liver+tissue»
Term Criterion 3
K2 : «"liver tissue AND cell phenotype"»
Evidence 3
Search result webpage : «https://www.ncbi.nlm.nih.gov/pubmed/advanc ed»
Decision Criteria
Articles: articles are selected based on how the trials are made and, overall, on how the study was made. E.g. how many people were involved in the trial, if there was a control group
Fact : decision based on the following filters
"article types"
"meta analysis"
"Free full text"
5 years range on "Publication dates"
Term Criterion Labels
K1 = General term
K2 = Specific term related to K1
K3 = Specific term not related to K1
K4 = Term not related to subject

Figure 8: Decision Diagram (Diagram 5).

As the process progresses, there is a change about the feeling involved, e.g. from uncertainty to certainty, and from desire, e.g. information need and term refinement, to *de facto* intention, e.g. term definition, query formulation, results check and information extraction. This change is demonstrated by the evolution from the desire shown in Figure 4, 'find scientific articles on the subject liver tissue', to the data object extracted shown in Figure 7.

2.4. Method

The first research cycle built upon the Design Science Research principles, also known as DSR, intended to model searches performed through Web search tools or portals of educational resources. The purpose, as a research goal, was to understand users' search and decision processes, their intentions and Web search patterns and to use KiP to model this knowledge process prompted by users' information gathered search activities.

The Design Science Research epistemology is motivated by the desire to improve the environment by introducing new and innovative artifacts[7], as well as the processes used to build these artifacts, as signaled by Simon [63]. A desire similar to the one driving this part of the study, focused on mapping searching processes – the process in which the artifact is built, using Web search engines or search systems – the environment to be improved, applying the Exploratory Search KiP model – the introduced artifact.

A problem-solving artifact designed from specific contexts and evaluated under these same contexts throughout a series of iterations, is one of the major features of Design Science Research. As expressed by Pimentel [55], an artifact is designed from assumptions about how people interact, how organizations work, and how their practices and culture influence a given context. By trying to explain the theoretical conjectures that

[7] Artifacts in DSR can broadly include models, methods, constructs, instantiations, algorithms and theories.

guide the artifact's design, the designer theorize about these presumptions in terms of problems and solutions and as a result, create new knowledge. As a first step to achieve a broader goal, e.g. identification of query reformulations, our artifact-model intents to help in the visualization of Web search patterns. By clarifying the patterns, the best practices associated to users' decision-making process can be used to understand Web searches.

One of the main criticisms that Design Science Research as a scientific method incites is if design – in the form of an artifact – can be considered a research. Owen [47] argues that "knowledge is generated and accumulated through act" in an iterative process, shown as a feedback cycle, in which the researcher's creativity and background knowledge is used to create an experiment that is not naturally present but occurs as a result of the preparative or investigative procedure. The experiment generates an artifact, which in turn is evaluated to build more knowledge. Vaishnavi, Kuechler and Petter [70] define it as being primarily "research using design as a research method or technique", almost as if the learning was acquired through building. It is a proposition not new in areas such as education and health care, where treatments and curricula are designed and empirically evaluated [70].

Another valid criticism posed to the use of design as a research method concerns the differences between the intellectual property created by its use and the state-of-practice knowledge derived from the work of designers, consultants and Research and Development personnel throughout the world. In this regard, Vaishnavi, Kuechler and Petter [70] point out a difference called the "production of interesting (to a community) new and true knowledge". More than a semantic effort, the statement provides a way to differentiate a proprietary knowledge from an academic one – which could become more intertwined as our society intensifies the shift from an industrial to a knowledge-based one – as both types of knowledge can be drawn from scientific endeavors.

The balance between the design project success and the knowledge learned by the greater community can summarize the difference. What is meant by knowledge learned is the transfer of new knowledge from the originator to a secondary user – i.e. the broadness in which the expertise involved and the knowledge derived are shared. Takeda *et al.* [66] proposed a design process model, which helps understand swiftly a typical DSR effort. It suggests five main steps that can also be interpreted as a knowledge flow: awareness of problem, suggestion, development, evaluation, and conclusion.

Awareness of problem is characterized by a conscious aware of an effort needed to solve a problem or meet a necessity. In the case of a research problem, it may come from multiple sources but it is similar in its role as an opportunity provider of new findings to be applied at the researcher’s field. In our current research, we set to map the search activities that influence on user’s search success and their impact in help tuning the user's mental image regarding the subject sought. As a lack of information to design a solution or to define the problem accurately taunts users search success, their previous expertise in searching can narrow the gap between the information retrieved by a Web search engine and its match to user’s intention. If query reformulations could be successful identified and classified, this knowledge could be used to build information systems, e.g. Web search engines, capable not only to retrieve relevant information but also to understand a Web search pattern and adjust the results accordingly. As a result, the search process could be accelerated – which could consequently – contributes to the user’s learning process while searching.

Suggestion, as a DSR step, can be summarized as the proposal developed based on the awareness of a problem. It is more accurately defined as an idea for the problem solution, which will be tested and modified as the research develops. In this research, to view exploratory search as a Knowledge-intensive Process was a challenge to approach the subject differently than the results ranking improvement efforts that typically characterized Information Retrieval field of study. It created a necessity to model this new vision in order to visualize the activities involved and its interactions, hence prompting the proposition of a new model to exploratory search.

Development is the tentative design further advanced and implemented as an artifact throughout the research. Whereas evaluation, is the step in which the artifact is assessed according to criteria usually defined in the suggestion phase. The first phase of the present research was comprised of one Design Science Research cycle and two evaluations – one using a Case Study scenario and other applying the model to search log analysis – both will be properly described further on in this Chapter.

Conclusion is the step in which the research results are consolidated and the knowledge gained categorized. In Design Science Research, as in any scientific method, the conclusion appropriately reports the research effort and its knowledge contribution. As the depth of knowledge contribution outputs can vary, the thesis research focuses on:

1) Report the Exploratory Search KiP model and how it contributes to the state-of-the-art;
2) Report the inferences regarding users search behavior, their decision-making process while searching online, and the shortcuts they use towards understanding.

2.4.1. DSR cycles to develop the Exploratory Search KiP model

Artifact abstraction, as commented by Gregor and Hevner [21], may be one of DSR main contributions. The term should be understood as a review to Diderot´s entry of "art" in his famous eighteenth century *Encyclopédie* and in which he stated: "every 'art' [technique] has its speculative and its practical side". Diderot's concept of speculation regarded the theoretical knowledge of the principles that embodies the technique. The concept of practice, the cited "practical side", in its turn, regarded the application of these principles.

By using the term artifact, this research intends to refer Goldkuhl [20] and considers the terminology as "something that has or can be transformed into an artificially made object or process". Information Technology artifacts have as a feature some degree of abstraction readily converted to a material existence, such as an algorithm that is converted to a software. A theory – as understood by Diderot's concept of speculation – is more abstract, with a nonmaterial existence, and contains knowledge to justify and rationalize the concepts behind a materially existing artifact.

The embodiment of three closely related cycles of activities presented by Hevner [22] can reconcile the apparent inconsistencies that a term such as "artifact abstraction" may provoke. These three cycles – Relevance Cycle, Rigor Cycle and Design Cycle – recognize the importance of both contributions (artifact and abstraction) made in the form of viable objects that contribute at more conceptual levels. The Relevance Cycle requires that the contextual environment's conditions be inputted into the research and that the research artifacts be introduced into environmental field-testing. The Rigor Cycle, for its turn, provides grounding theories, methods that go along with domain experience and expertise, and adds the new knowledge generated by the research to a growing knowledge base. Finally, the Design Cycle is central to support an iterative research activity for the construction and evaluation of design artifacts and processes.

The three cycles framework overlays three inherent research cycles:

a) The bridging of the contextual environment from which the research problem emerged to the research project activities;
b) The connection of the research project activities to the knowledge base of scientific foundations, experience, and expertise that support the research as a whole;
c) The characteristic act of repetition between the core scientific research activities of making conjectures, deriving predictions or logical consequences, and carrying out experiments and empirical observations to evaluate the derived predictions or consequences.

The first cycle of the present research sought to blend KiP and KIPO perspectives into search activities to better represent the characteristics and behaviors involved. The modeling through KIPN diagrams helped to deep dive into Relevance Cycle to propose the Exploratory Search KiP model, with the continuous use of the Rigor Cycle to review the existing knowledge regarding exploratory search and information seeking and of the Design Cycle to improve the model. The search activities, characteristics and behaviors were based on concepts supported by Vakkari [73], White and Roth [75], Wildemuth and Freund [76], Marchionini [36] and Ellis *et al.* [14]. Those related to exploratory browsing, focused searching, search query formulation and reformulation, and search behavior attributes were used to ground the artifact abstraction into the proposed Exploratory

Search KiP model.

The contribution of Exploratory Search KiP model resides in presenting not only the information-gathering behavior embedded in an exploratory search, but also the knowledge process prompted by the information gathered since that information helps forming an evolving mental image of the searched subject. The model can be successfully applied to Web searches to help identifying Web search patterns and best practices associated to decision-making process.

2.4.2. Case Study Scenario

The Exploratory Search KiP model was tested and revised in a collaboration between Italian, German and Brazilian Universities, using data from a previous study conducted by Bortoluzzi and Marenzi [5]. In this Case Study, six expert teachers from different Italian school levels were selected and observed while searching online resources on the Web through think-aloud protocols and interviews in which they commented on their search habits and choices. Videos showing the teachers' interactions with search tools along with think-aloud protocol served as a basis for analyzing the model's effectiveness in representing the knowledge process of these search interactions. Figures 9 and 10 show the think-aloud procedure's dynamics in one of the teachers' original searches.

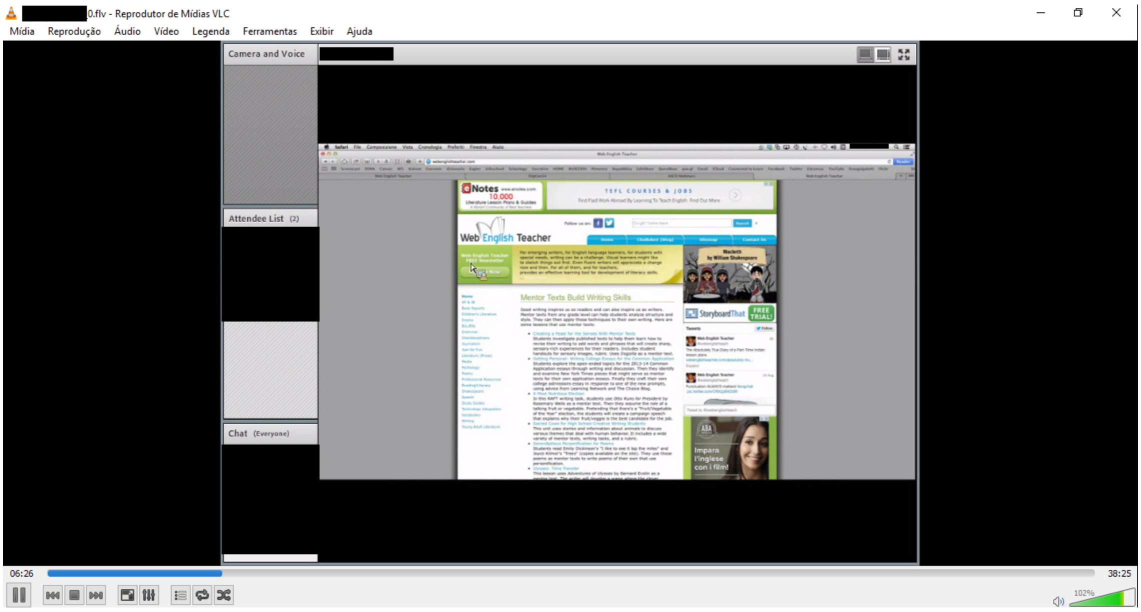

[redacted]0.flv - Reprodutor de Mídias VLC
Mídia Reprodução Áudio Vídeo Legenda Ferramentas Exibir Ajuda
Camera and Voice [redacted]
Attendee List (2)
Chat (Everyone)
Web English Teacher
Mentor Texts Build Writing Skills
StoryboardThat
06:26
38:25
102%

Figure 9: The teacher described the activity performed.

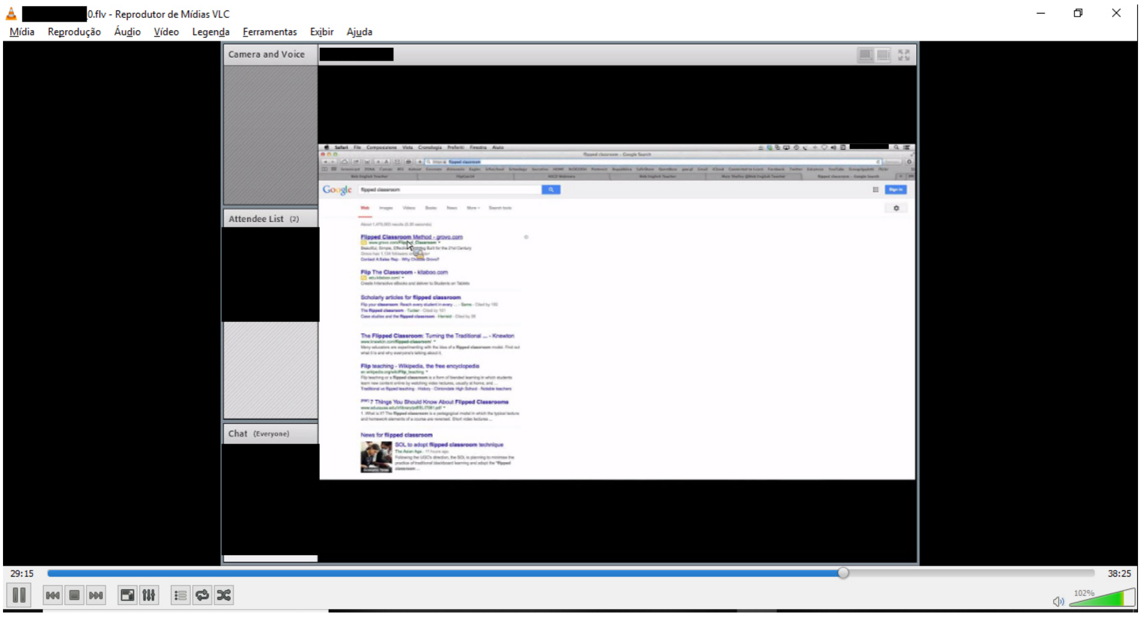
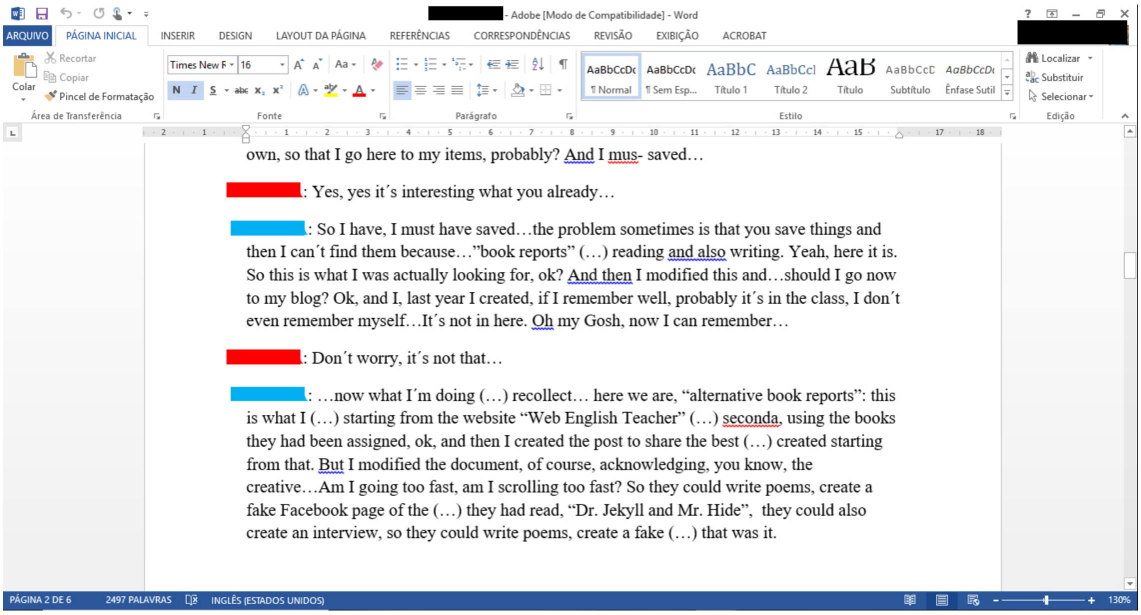

[redacted]0.flv - Reprodutor de Mídias VLC
Mídia Reprodução Áudio Vídeo Legenda Ferramentas Exibir Ajuda
Camera and Voice [redacted]
Attendee List (2)
Chat (Everyone)
Google flipped classroom
Flipped Classroom Method - growo.com
Flip The Classroom - kitaboo.com
Scholarly articles for flipped classroom
The Flipped Classroom: Turning the Traditional ... - Knewton
Flip teaching - Wikipedia, the free encyclopedia
7 Things You Should Know About Flipped Classrooms
News for flipped classroom
29:15
38:25
102%

Figure 10: The teacher explained why had decided to click in a particular link.

To perform the searches the teachers were presented with three tasks. The first one was to open a portal they usually used to find teaching resources and repeat a search they had previously performed and considered a successful one, i.e. a teaching resource found that could be used in class. The goal was to understand what was considered a successful search and the steps taken to achieve it. As for the second task, they were asked to return to the same portal to look for a new resource or topic they had not investigated yet, but they would be interested in using. The intention of this particular task was to verify the occurrence of a Web search pattern that could be responsible for the search's success. Finally, the third task was to look for a completely new resource, starting the new search through a Web search engine. This task's objective was to observe the entire search process and the application of the steps and patterns used in both previous tasks.

The transcription of the searches (Figure 11) allowed a detailed analysis regarding the reasons for the teachers' behaviors during the searches and provided insights into his/her motivation while performing the activities.



own, so that I go here to my items, probably? And I mus- saved…

[redacted]: Yes, yes it´s interesting what you already…

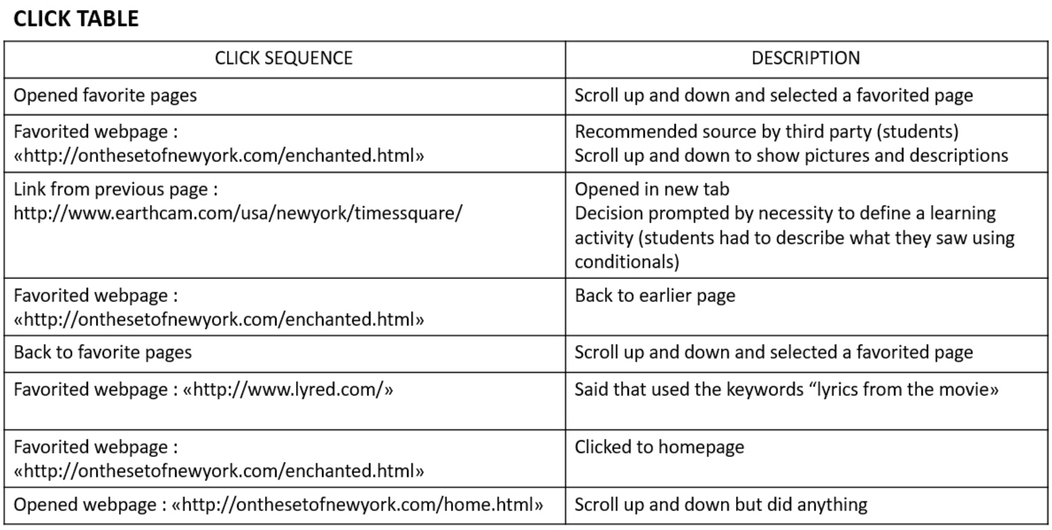
[redacted]: So I have, I must have saved…the problem sometimes is that you save things and then I can´t find them because…"book reports" (…) reading and also writing. Yeah, here it is. So this is what I was actually looking for, ok? And then I modified this and…should I go now to my blog? Ok, and I, last year I created, if I remember well, probably it´s in the class, I don´t even remember myself…It´s not in here. Oh my Gosh, now I can remember…

[redacted]: Don´t worry, it´s not that…

[redacted]: …now what I´m doing (…) recollect… here we are, "alternative book reports": this is what I (…) starting from the website "Web English Teacher" (…) seconda, using the books they had been assigned, ok, and then I created the post to share the best (…) created starting from that. But I modified the document, of course, acknowledging, you know, the creative…Am I going too fast, am I scrolling too fast? So they could write poems, create a fake Facebook page of the (…) they had read, "Dr. Jekyll and Mr. Hide", they could also create an interview, so they could write poems, create a fake (…) that was it.

Figure 11: A transcription sample from one of the searches.

To map the search movements operationally, click tables were constructed to formalize and describe the sequences in which the teachers performed their actions (Figure 12).



CLICK TABLE

CLICK SEQUENCE	DESCRIPTION
Opened favorite pages	Scroll up and down and selected a favorited page
Favorited webpage : «http://onthesetofnewyork.com/enchanted.html»	Recommended source by third party (students) Scroll up and down to show pictures and descriptions
Link from previous page : http://www.earthcam.com/usa/newyork/timessquare/	Opened in new tab Decision prompted by necessity to define a learning activity (students had to describe what they saw using conditionals)
Favorited webpage : «http://onthesetofnewyork.com/enchanted.html»	Back to earlier page
Back to favorite pages	Scroll up and down and selected a favorited page
Favorited webpage : «http://www.lyred.com/»	Said that used the keywords "lyrics from the movie»
Favorited webpage : «http://onthesetofnewyork.com/enchanted.html»	Clicked to homepage
Opened webpage : «http://onthesetofnewyork.com/home.html»	Scroll up and down but did anything

Figure 12: A sample of a click table used to map the actions performed.

To explain the modeling process better, it is necessary to describe briefly the criteria used to identify what was called 'desire', which was the goal that motivated the search. It is also necessary additional clarification on what determined the 'desire' to be transformed into action – characterized in the model by the term 'intention' – and how each action was used to determine what was called a 'mental image', which was the representation of how the user's awareness about the subject searched evolved.

The teacher stated the 'desire' before starting the search, e.g. "I want to find scientific articles". At the moment the teacher decided to open a Web search engine or a portal and chose a search term to begin his/her search, it was considered that he/she had passed from desire to intention – e.g. the usage of the search term "liver tissues". The psychological 'state change' from desire to intention occurred when the teacher decided to click on a certain link during the results check.

To be able to consider a mental image's formation or refinement properly, three criteria were used. The first one, a subjective criterion, was identified when the results check created a questioning declared by the teacher, e.g. "which articles are available on this topic?". The second was observed when a refinement of the search term occurred, e.g. from “coaching with compassion” to “coaching with compassion scale OR questionnaire OR questionnaires OR questions”. The third, an objective criterion, was discerned when the teacher clearly stated his/her awareness of a situation or fact prompted by the search, e.g. "I understood this topic" or "this will help me". Generally, this third criterion was reinforced by information extraction, e.g., a file downloaded or a link marked as favorite.

How to map properly what the teachers did or tried to do during the search was an initial challenge encountered. Icons and visual images from KIPN helped but they were not enough. The teachers own explanations during the think-aloud protocols led to the inclusion of labels and comments to better capture and represent the teachers’ actions and words.

It was possible to see that the decisions involved in the activities were, in a general sense, repeated. They were used to stimulate terms selection and define what results would be examined, generating two types of decision criteria adopted by the teachers throughout the searches. It is noteworthy that the two criteria types would not be detectable without the visualization provided by the Exploratory Search KiP model. The first decision criteria concerns keywords and was named “Term Criterion Labels”. The labels were categorized into four kinds:

(K1) related to a general term or keyword usually used to start the search task;

(K2) regarding a more specific term related to K1;

(K3) also a more specific term, but in this case not related to K1;

(K4) represents a term not related to the subject, but it rather indicates a completely new search triggered by the previous one.

The second type involved criteria related to the decision of what link to click and what resource to choose, and were simply named “Decision Criteria”. These criteria mirrored the teachers’ expertise on their knowledge domain and experience about their students’ personalities and class context.

Initial mappings encountered difficulties in capturing the relevance of context variables, i.e. what type of resource the user is looking for and for what purpose, and personal expertise used by the teachers as search filters. Figure 13 shows an example of this difficulty characterized by the failed attempt to represent properly a mental image about the subject – a lack of clarity regarding the type of refinement derived from the results check – and by the poorly identification of the search activities involved, only two from the four originally proposed.

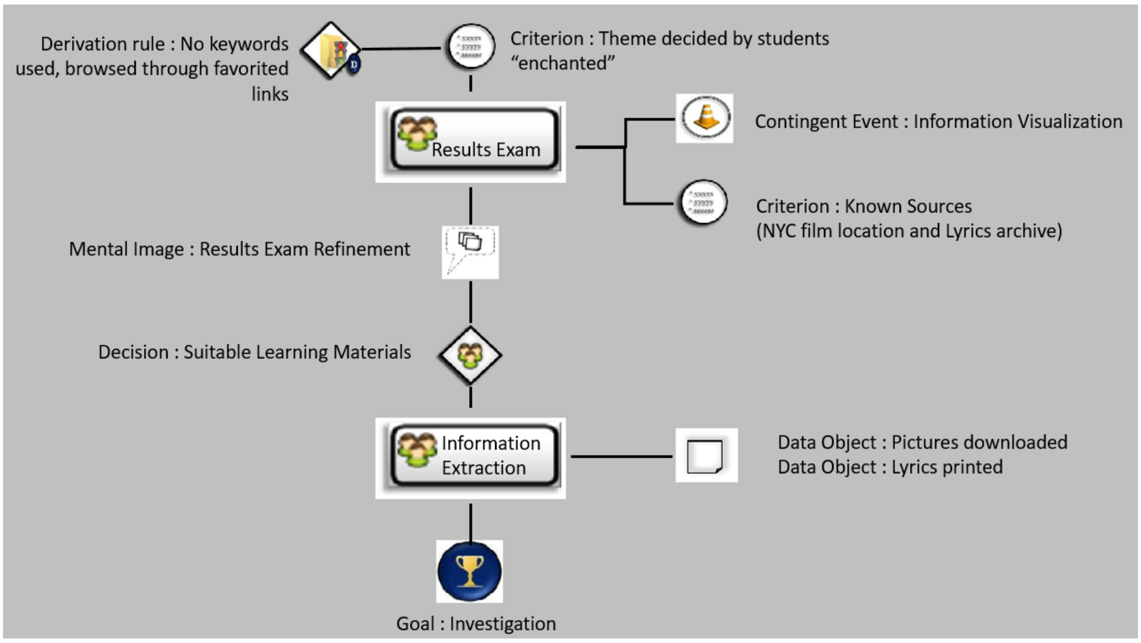
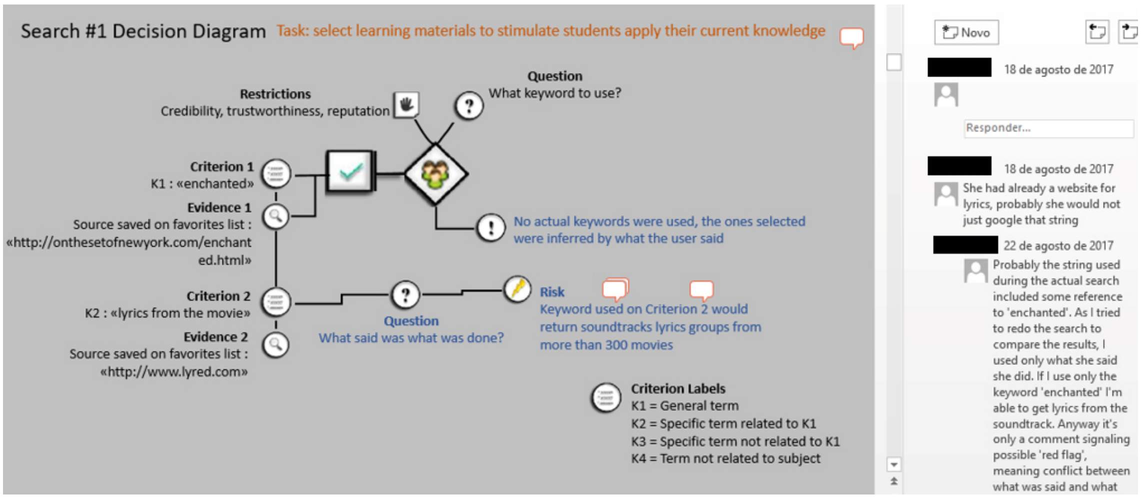

Derivation rule : No keywords used, browsed through favorited links
Criterion : Theme decided by students "enchanted"
Results Exam
Contingent Event : Information Visualization
Criterion : Known Sources (NYC film location and Lyrics archive)
Mental Image : Results Exam Refinement
Decision : Suitable Learning Materials
Information Extraction
Data Object : Pictures downloaded
Data Object : Lyrics printed
Goal : Investigation

Figure 13: Initial modeling of one of the searches based on video analysis.

These difficulties were mitigated through validation interactions and review processes based on user feedback conducted in association with University of Udine and L3S Research Center at Leibniz University of Hannover. Every mapping was "dissected" in order to signal possible "red flags", meaning conflicts between what the user said and what was really done during the search. The persons involved were then asked to provide more information about the particular search. The same happened when there were no clear evidence about a mental image representation. This process helped to refine the Diagrams. Figure 14 shows an example of such interactions.



Search #1 Decision Diagram Task: select learning materials to stimulate students apply their current knowledge

Restrictions
Credibility, trustworthiness, reputation

Question
What keyword to use?

Criterion 1
K1 : «enchanted»

Evidence 1
Source saved on favorites list : «http://onthesetofnewyork.com/enchanted.html»

No actual keywords were used, the ones selected were inferred by what the user said

Criterion 2
K2 : «lyrics from the movie»

Question
What said was what was done?

Risk
Keyword used on Criterion 2 would return soundtracks lyrics groups from more than 300 movies

Evidence 2
Source saved on favorites list : «http://www.lyred.com»

Criterion Labels
K1 = General term
K2 = Specific term related to K1
K3 = Specific term not related to K1
K4 = Term not related to subject

[redacted] 18 de agosto de 2017

[redacted] 18 de agosto de 2017
She had already a website for lyrics, probably she would not just google that string

[redacted] 22 de agosto de 2017
Probably the string used during the actual search included some reference to 'enchanted'. As I tried to redo the search to compare the results, I used only what she said she did. If I use only the keyword 'enchanted' I'm able to get lyrics from the soundtrack. Anyway it's only a comment signaling possible 'red flag', meaning conflict between what was said and what

Figure 14: Example of a review process conducted at one of the searches mapped.

2.4.3. Exploratory Search KiP model as a Guide to Understand Logs

By visualizing how the purpose that prompted the search became intentions and how intentions drove the search activities, the model also made visually possible to observe the formation and refinement of a mental image regarding the subject sought. Mental image, defined in the context of this thesis as an understanding about the subject capable of empowering rational arguments and elevating one's awareness, also helps to understand how the search intention is modified during search activities.

It performs this particular task by allowing an observer to get a fair picture of how users' decision processes were related to the links they chose to click, the terms they included in their queries, and how this specification helped them to refine the search results. The Decision Diagram shown in Figure 8 is an example on how the visualization described in the last period works. Note the query reformulation types mapped by the Term Criterion Labels led to webpages visited, which were characterized as evidences. On the third query reformulation type mapped, "liver tissue AND cell phenotype", filters were used to tune the results. Those filters prompted the last 'intention' mapped on Figure 6, regarding a link examination and the 'Mental Image' shown in Figure 4, regarding the search refinement.

To investigate if applying the model to data at a less comprehensive level of context than the one supplied by the think-aloud protocol data would provide the possibility to generalize the model; it was decided to use the Exploratory Search KiP model's elements to analyze search logs. The purpose was to verify its applicability in searches with less specific situations than those found by the teachers. The logs were collected from an online professional community[8] of student teachers and expert teachers during a workshop aimed to analyze how teachers search for and tag online resources for their teaching practice. During a nine-day activity, 22 users performed 1,249 search sessions.

2.5. Evaluation Setup

From the data of 52 activities stored in transaction logs of this online teacher professional community, it was mined the ones related to exploratory search activities as presented in Exploratory Search KiP model. In order to do that, not only the activity description but also the parameter involved were selected, e.g., "searching" activity implied the parameter "query". The correlation is described in Table 1.

[8] A platform named LearnWeb-OER, which purpose is to enhance collaborative searching and sharing of educational resources.

Table 1: Logs and Exploratory Search KiP model activities correlation.

Activities from the logs	Exploratory Search KiP model activities	Additional explanation
“searching” and “group resource searching”	Corresponding to “search term selection” and “query formulation” activities	Represent the formulation of an initial query and subsequent reformulations
“opening resource”	Corresponding to “results check” activity	Represent the action of clicking and opening a particular resource
“downloading”	Corresponding to “information extraction” activity	Represent the action of extracting a particular resource
“tagging resource”	Corresponding to “information extraction” activity	Represent the action of saving a particular resource and a measurement of understanding. If the user can assign tags, this action shows understanding the subject searched and the resource selected

Figure 15 shows a sample taken from one of the sessions performed by User 11270 and modeled as an Exploratory Search KiP. It is possible to observe the specialization of the search term from a general one, "metaphor", to a more specific, "metaphors for learning" and "metaphors for language teaching". For comparison purposes, Figure 16 shows the sequence of actions captured by the logs for the session.

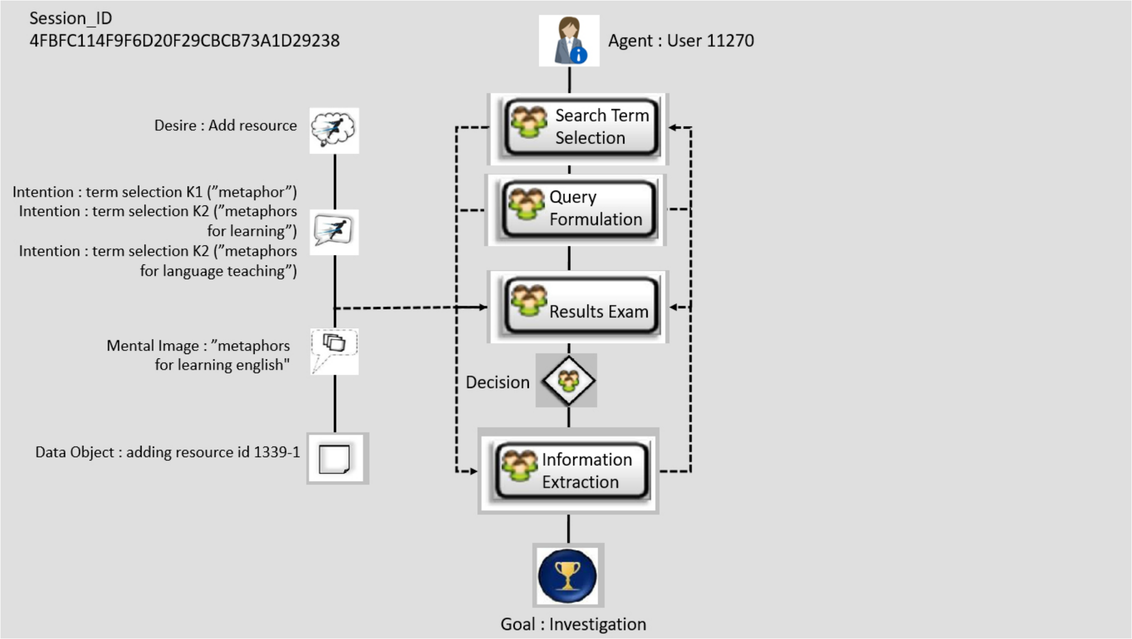
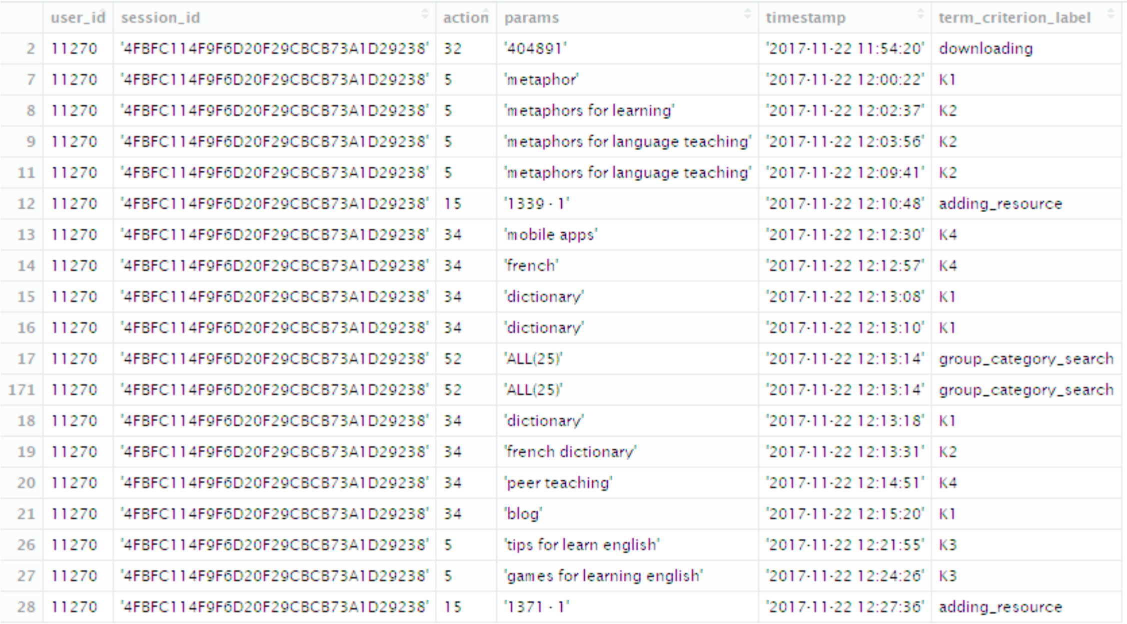

Session_ID
4FBFC114F9F6D20F29CBCB73A1D29238
Agent : User 11270
Desire : Add resource
Search Term Selection
Intention : term selection K1 ("metaphor")
Intention : term selection K2 ("metaphors for learning")
Intention : term selection K2 ("metaphors for language teaching")
Query Formulation
Results Exam
Mental Image : "metaphors for learning english"
Decision
Data Object : adding resource id 1339-1
Information Extraction
Goal : Investigation

Figure 15: Sample of an Exploratory Search KiP modeled for one of the sessions.



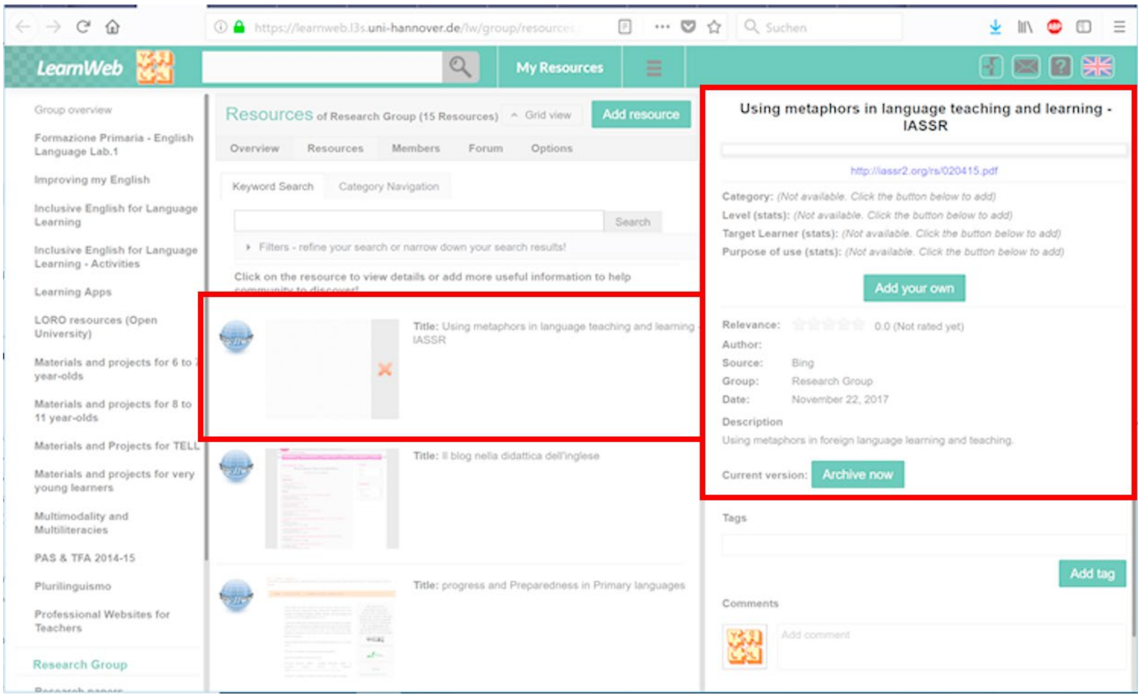
	user_id	session_id	action	params	timestamp	term_criterion_label
2	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	32	'404891'	'2017-11-22 11:54:20'	downloading
7	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	5	'metaphor'	'2017-11-22 12:00:22'	K1
8	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	5	'metaphors for learning'	'2017-11-22 12:02:37'	K2
9	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	5	'metaphors for language teaching'	'2017-11-22 12:03:56'	K2
11	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	5	'metaphors for language teaching'	'2017-11-22 12:09:41'	K2
12	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	15	'1339 - 1'	'2017-11-22 12:10:48'	adding_resource
13	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	34	'mobile apps'	'2017-11-22 12:12:30'	K4
14	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	34	'french'	'2017-11-22 12:12:57'	K4
15	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	34	'dictionary'	'2017-11-22 12:13:08'	K1
16	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	34	'dictionary'	'2017-11-22 12:13:10'	K1
17	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	52	'ALL(25)'	'2017-11-22 12:13:14'	group_category_search
171	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	52	'ALL(25)'	'2017-11-22 12:13:14'	group_category_search
18	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	34	'dictionary'	'2017-11-22 12:13:18'	K1
19	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	34	'french dictionary'	'2017-11-22 12:13:31'	K2
20	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	34	'peer teaching'	'2017-11-22 12:14:51'	K4
21	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	34	'blog'	'2017-11-22 12:15:20'	K1
26	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	5	'tips for learn english'	'2017-11-22 12:21:55'	K3
27	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	5	'games for learning english'	'2017-11-22 12:24:26'	K3
28	11270	'4FBFC114F9F6D20F29CBCB73A1D29238'	15	'1371 - 1'	'2017-11-22 12:27:36'	adding_resource

Figure 16: Logs from the session listed by timestamp.

2.6. Results

The pattern captured by the modeling corresponds to the timestamp "2017-11-22 12:00:22" through "2017-11-22 12:10:48". In this space of 10 minutes, it is possible to identify the formation of a mental image based on the terms “metaphors for learning” and “metaphors for language teaching”, which resulted in the addition of a resource represented by the extraction of the data object "resource id 1339-1", shown in Figure 17.

Back to Figure 15, it was possible to observe how the purpose that drives the search – find a resource available in this online professional community and add it to the user’s repository – turned into real actions, the definition and refinement of terms from “metaphor” to “metaphors for language teaching”.


https://learnweb.l3s.uni-hannover.de/lw/group/resources
Suchen
LearnWeb
My Resources
Group overview
Formazione Primaria - English Language Lab.1
Improving my English
Inclusive English for Language Learning
Inclusive English for Language Learning - Activities
Learning Apps
LORO resources (Open University)
Materials and projects for 6 to 7 year-olds
Materials and projects for 8 to 11 year-olds
Materials and Projects for TELL
Materials and projects for very young learners
Multimodality and Multiliteracies
PAS & TFA 2014-15
Plurilinguismo
Professional Websites for Teachers
Research Group
Resources of Research Group (15 Resources)
Grid view
Add resource
Overview
Resources
Members
Forum
Options
Keyword Search
Category Navigation
Search
Filters - refine your search or narrow down your search results!
Add your own
Archive now
Tags
Add tag
Comments
Add comment

Click on the resource to view details or add more useful information to help community to discover!

Title: Using metaphors in language teaching and learning - IASSR

Title: Il blog nella didattica dell'inglese

Title: progress and Preparedness in Primary languages

Using metaphors in language teaching and learning - IASSR

http://iassr2.org/rs/020415.pdf

Category: (Not available. Click the button below to add)
Level (stats): (Not available. Click the button below to add)
Target Learner (stats): (Not available. Click the button below to add)
Purpose of use (stats): (Not available. Click the button below to add)

Relevance: ☆☆☆☆☆ 0.0 (Not rated yet)
Author:
Source: Bing
Group: Research Group
Date: November 22, 2017

Description
Using metaphors in foreign language learning and teaching.

Current version:

Figure 17: The resource added to the online professional community.

The activities “searching” and “group resource search” from the logs were related to Exploratory Search KiP model activities “Search Term Selection” and “Query Formulation” because they displayed patterns and transitions of query reformulation. Both activities exhibited the parameter “query” and possessed keywords or more structured queries that could be mapped using the Term Criterion Labels presented in the Exploratory Search KiP model’s Decision Diagram at subsection 2.4.2.

"Opening resources" was related to the "Results Check" activity based on its link-click property and the log's "downloading" and "tagging resources" activities to "Information Extraction" as they are connected to actions of extraction or saving the data considered suitable.

The analysis was performed using R Programming to stabilize the data and Google Charts to compare users search activities to Exploratory Search KiP model activities. By using the Term Criterion Labels to map the search pattern such as the example presented in Figure 18, it was possible to observe a search profile displayed by the user to draw inferences and shortcuts to support his/her understanding. In this case, a refinement from general term (K1) to a specific one (K2), followed by a change of subject (K4) until his/her decision to use a more specific subject not related to the general term previously used (K3), but better suited to put into action his/her intention. A move implying a broader awareness about the subject than before.

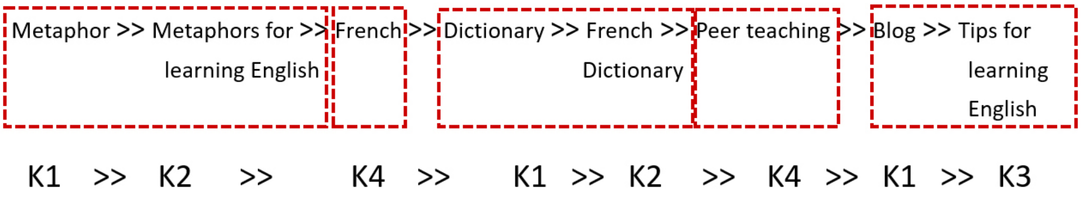

Metaphor >> Metaphors for >> learning English
French >>
Dictionary >> French >> Dictionary
Peer teaching >>
Blog >> Tips for learning English
K1 >> K2 >> K4 >> K1 >> K2 >> K4 >> K1 >> K3

Figure 18: Example of Term Criterion Labels being used to map the search pattern.

The overall percentages of actions, calculated by adding the frequency of each action performed by users in every session, show that the "Results Check" activity, i.e. opening resource, from Exploratory Search KiP Model is applied only to a certain extent to enable users to select sources by quality and relevance and extract the resources, i.e. downloading, as Figure 19 shows.

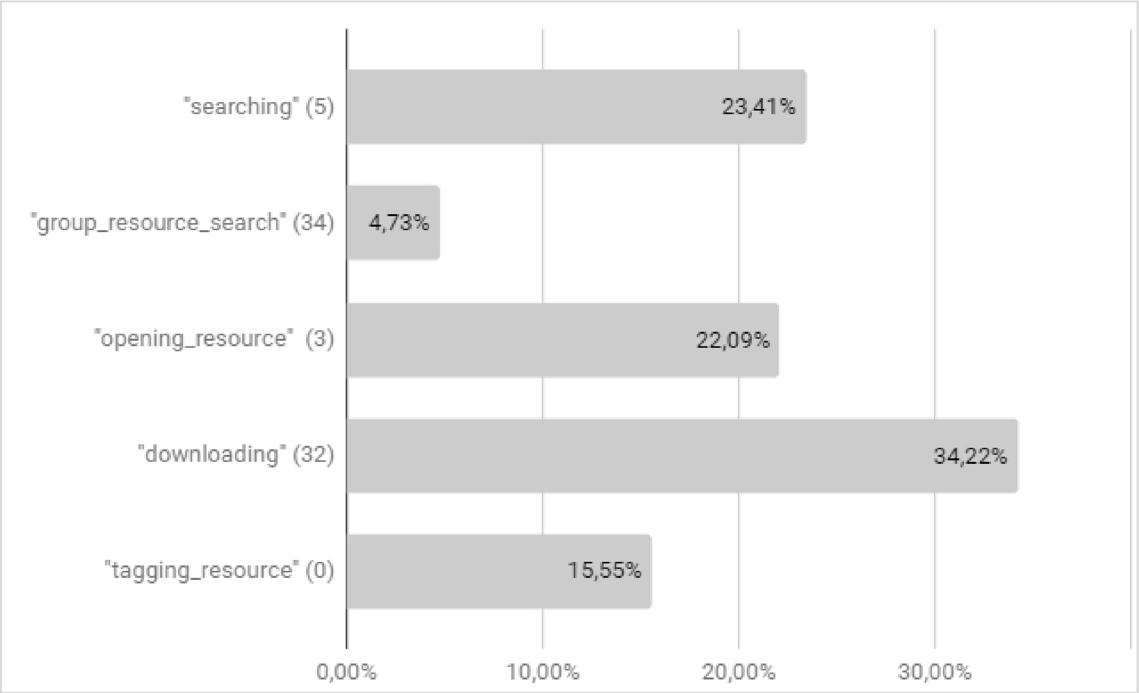


Figure 19: Overall percentage of actions performed by users.

The most performed actions were those connected to extraction and saving activities, i.e., downloading and tagging resource, representing almost half of them (49.77%). Query related activities, i.e., searching and group resource search, correspond to the second most performed action (28.14%) and it is noticeable that although important, the act of checking the search results, i.e., opening resource, is not a decisive factor to determine suitability. Perhaps due to this particular search environment's characteristics, an online professional community, and differences of its users – student teachers and expert teachers – regarding subject domain level, e.g., less and in-depth level of context knowledge, and search circumstances, e.g., less specific situations.

2.7. Conclusion

Although correlating the action of opening resources to the action of downloading them seems natural, observing the users individually shows a different story. It is possible to notice that the frequency of resources opened by users during their sessions indicates that the "Results Check" activity is not performed each time information is retrieved during a search. In addition, the frequency of downloading resources indicates that users did not necessary examine a result through link-click to decide if a data object is worth retrieving. As shown in Figure 20, two of the three users downloaded resources without opening them previously. In these cases, the users' decision criteria seem to be more deeply internalized or tacit to the point that checking the content becomes unnecessary.

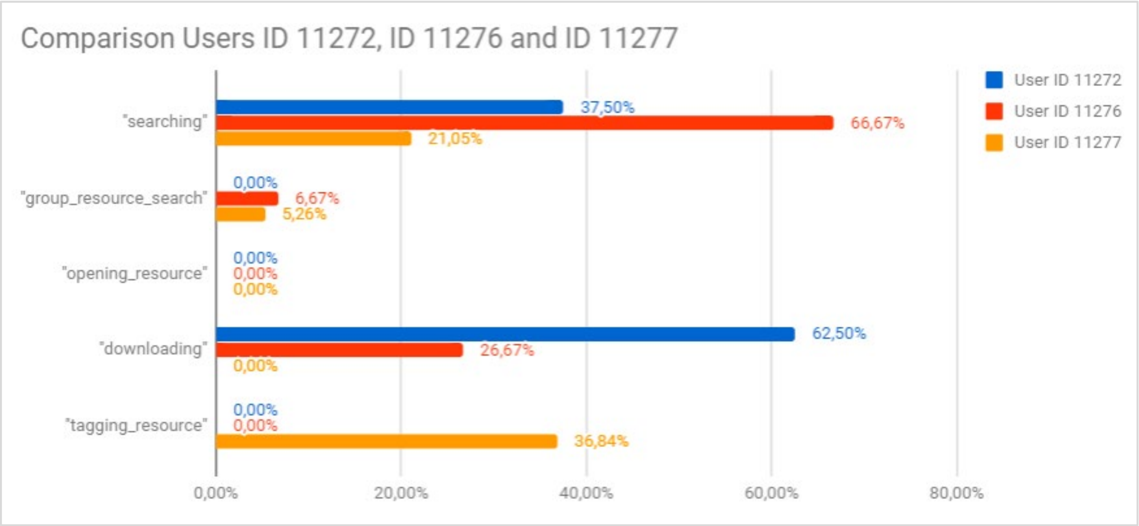


Figure 20: Percentage of actions performed by users 11272, 11276 and 11277.

Performing a series of scatter plots drawn from our analysis, we were able to see the frequency of the actions performed by users and distributed through the sessions (Figure 21).

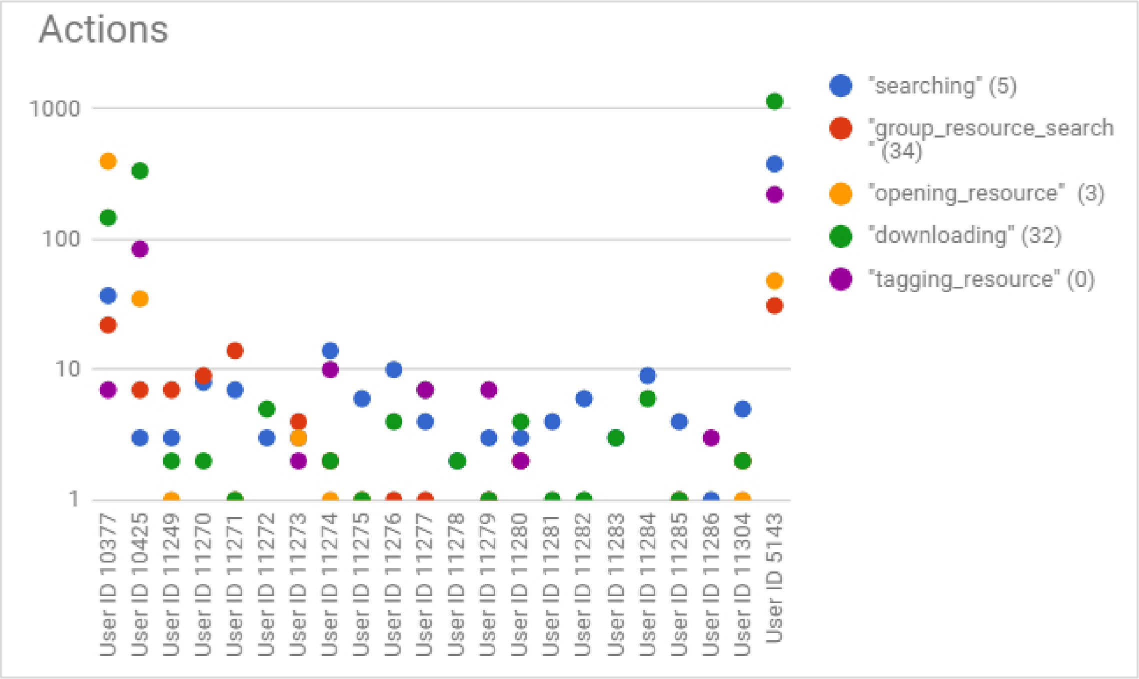


Figure 21: Actions performed by users distributed through the sessions.

Even though not every user performed each action, the frequency of those actions performed indicates an Exploratory Search KiP pattern. The pattern is demonstrated by showing a synchronicity of actions and a high number of query reformulation, as most of the blue and red dots – representing query related activities – are greater than five. Both alongside an average of time spent of 12.89 minutes per session seems to indicate that the users spent a fair amount of time in most of their sessions.

3. QUERY STATES AND TRANSITIONS BASED ON EXPLORATORY SEARCH AS A KNOWNLEDGE-INTENSIVE PROCESS

3.1. Introduction

As Web search engines are one of the most frequently used tools to reach information on the Web [48, 50], analyze Web search patterns became a frequent issue while investigating user interaction with Information Retrieval systems, especially within their information-seeking and topic-classification behaviors during their search for information [24, 48, 64, 72]. In examining the searching process, the usage of the concept of query states to model the sequence of human-machine interaction during a search [7, 25] has been a favorite choice. The state of a query is based on the set of terms used in consecutive queries within a session and is determined by classes of behavior, such as uniqueness, generalization, specialization and reformulation.

Historically, researchers typically identify user actions on an information searching system, classify these actions into states and then build a state map or matrix of possible moves [25, 53, 54, 57]. The state-based research of search pattern has two main approaches: Stochastic Process and Markovian. The Stochastic Process approach characterizes pattern as a sequence of state changes and considers that in order to arrive at a certain state, the preceding states are important. It focuses on search actions and system responses. The Markovian approach assumes that future states depend only on the current state, not on past events. Therefore, it uses Markov models to control the variety of variables and examines search modifications in search terms and feedback states. It focuses on query reformulation.

Query reformulation tends to be prompted by several reasons such as find a specific information, verify a specific fact or find new information, rather by a particular one [16]. Although specialization is the most frequent information reason to reformulate a query, generalization – e.g. broadening a query by reducing the words in it – is also used, especially to find relevant information not retrieved by the use of specific terms and to bring new information. Sarigil *et al.* [15], describing their findings about query length, determined that most Web queries[9] contains one to three terms and as more terms are included (e.g. n > 10), the ability to retrieve relevant information starts to decrease as becomes harder to match the query to a document that contains all query terms.

Rha, Shi and Belkin [16] asserts that in order to identify users' reasons causing query reformulation, it is also necessary to understand users' previous situations. In an approach closely in concept to Stochastic Process' since purposes or goals of query reformulation – in the present thesis was opted to use the terms desires and intentions to relate the same meanings – can be affected by preceding events or their cognitive status in the interactive information seeking processes. Although the present study does agree that some previous intentions may lead users to have a particular purpose in modifying a query and induce them to add, remove or modify the words used, it is due to user's

[9] Query entries on search engines, portals and websites mainly.

satisfaction or not with the current query's retrieved data one of the main drivers to incite a query reformulation. More connected, in general notion, to Markovian approach.

The Markovian approach is also more adequate to represent the decision-making pattern disclosed by the "Term Criterion Labels" presented at Exploratory Search KiP model's Decision Diagram [35, 67], in which users change keywords and terms on their current queries – Query Formulation activity – to better tune the results retrieved by the search engine as seen in subsection 2.4.2. and section 2.6.

3.2. Related Work

The Markovian approach is characterized by the examination of search modifications in search terms and feedback states [64] with the usage of n-order Markov process to provide a best description of the data. Focused primarily on search or browsing actions and system responses, Jansen, Booth and Spink [25] cite research studies using Markov process for four main reasons:

1) To predict the users' future page requests or user query reformulation patterns;
2) To draw inferences of searcher topic interests by using visited uniform resource locators (URLs);
3) To predict future page requests using user's history and frequency of access;
4) As an effective design technique for a recommender system – combining with implicit search for that matter.

Researches focusing on the state transitions as users reformulate their queries during a session tend to rely on query reformulation. Rieh and Xie [60] conducting a qualitative analysis, reported three facets of query reformulation – content, format, and resource – with subfacets related to each of the three areas. Özmutlu and Çavdur [48] investigated the use of neural networks to automatically identify topic changes of queries within sessions and defined a terminology and sub-groups that echoed in subsequent studies [16, 25, 32]. They defined 'topic' as a group of queries submitted by a single user on a single topic and 'session' as a group of queries submitted by a single user.

Jansen, Booth and Spink [25] later updated the definition of session to "a series of queries submitted by a user and related interactions during an episode of interaction between the user and the Web search engine around a single topic", adding to it a temporal character.

Özmutlu and Çavdur [48], believing that queries showing terms modification was better understood by behavior classification, proposed the sub-groups generalization, specialization and reformulation to provide more insight to Web user search behavior analysis. The three sub-groups became the Web search patterns' core and were detailed by the authors as follow:

- ✓ Generalization: "The second query has fewer terms than the first query, and all terms of the second query are in the first query";
- ✓ Specialization: "The second query has more terms than the first query and includes all the terms of the first query";
- ✓ Reformulation: "Some (not all) of the terms of the second query are also included in the first query but the first query has some other terms that are not included in the second query".

Note the recurring mention to first and second queries, referring the modified state to the memoryless property of a Markov model.

Liu *et al.* [32] created a taxonomy of query reformulation with five categories: Generalization (G), Specialization (S), Repeat (RP), Word Substitution (WS) and New (N). The authors depicted Özmutlu and Çavdur's first and second queries [48] as Qi and Qi+1 respectively, being Qi+1 the query immediately following the query Qi in the same session. The five categories represent reformulation types and were described as follow:

- ✓ G: Qi and Qi+1 contain at least one term in common; Qi+1 contains fewer terms than Qi.
- ✓ S: Qi and Qi+1 contain at least one term in common; Qi+1 contains more terms than Qi.
- ✓ RP: Qi and Qi+1 contain exactly the same terms, but the format of these terms may be different.
- ✓ WR: Qi and Qi+1 contain at least one term in common; Qi+1 has the same length as Qi, but contains some terms that are not in Qi.
- ✓ N: Qi and Qi+1 do not contain any common terms.

As a Knowledge-intensive Process, Exploratory Search has some specificities of its own. The first of them, concerns Jansen, Booth and Spink's definition of session [25] as the "interactions during an episode of interaction between the user and the Web search engine around a single topic". The experience provided by analyzing search's interaction from both the think-aloud protocol and log analysis [35, 67] indicate that a user can – and often does – change the topic of a search during a session. It might be due to the user's lack of information at the search problem or solution definition stage. In any case, the current thesis tends to consider Özmutlu and Çavdur's generic way of dealing with the definition of a session [48] more suitable to exploratory search's necessity than Jansen, Booth and Spink's [25], only adding the temporal character of their definition to it.

The second specificity regards the use of syntactically different but semantically related terms, a case not covered by Liu *et al.*'s taxonomy [32]. An example of it is the use of a general word as 'spyware' as Qi and a more specific one as 'trojan horse' as Qi+1. Another specificity concerns the definition of an initial search state as a reformulation type, which was not a necessity of previous studies about query reformulation [16, 25, 32, 48, 60] but an important one to the present research due to the appearance of a recurring situation: the return to an initial state after specializations. A demonstration of such case is the following search sequence: 'lake como' ->′lake como vacation season' -> ′lake como′->′lake como rental prices'. Considering the set Q of the previous sequence as (q1, q2, q3, q4), note that q3 represents a return to the initial state q1.

3.3. ESKiP Query States and its Transitions

The complexity of a query reformulation taxonomy is reflect by the roles of reasoning and human expression it has to represent. This diversity is *per se* a major challenge and it increases when adding important questions to user's educational need when considering a search engine a learning tool. Is query states and transitions connected to users search strategies? User search strategies could be used to identify search profiles? Search profiles could be used to set retrieval ranking?

To start to answer the questions, ESKiP Taxonomy demanded a number of formalizations necessary to proper represent the knowledge a query reformulation induces. The formalization required not only subjects composing the action of Web search such as query and session but also those concerned to user behavior. An example is the definition of search profiles, which although is not direct connected to query states is important to the information retrieval level used to optimize search engines' results.

The specificities cited previously led to the following formalizations of key concepts important to better understand the second research cycle:

- ✓ *Term:* a series of keywords or characters used within a search engine and separated by white space or other type of separator.
- ✓ *Query:* a string of terms submitted by a user, combining all the texts, numbers and symbols entered into a search engine.
- ✓ *Session:* a group of queries submitted by a user during an event or a group of events of interaction between the user and the search engine around a single or multiple topics within a given period.
- ✓ *Exploratory Search Episode:* one or more sessions conducted by the same individual user within a given period.
- ✓ *Query Reformulation:* the process of altering a given query to improve search or retrieval performance.
- ✓ *Query States:* the set of terms used to reformulate queries within a session, determined by user's decision-making to include, exclude, modify or keep the set of terms. Also referred in this work as query reformulation types.
- ✓ *User Search Strategies:* relate to behavior attributes associated to user's decision-making that determines the way he or she decides to transition from query state to query state within a session or an exploratory search episode by including, excluding, modifying or keeping the chosen set of terms in order to refine the search engine's responses.
- ✓ *Search Profiles:* the set of user search strategies implemented that determines a Web search pattern.

One opportunity opened up by this thesis was the likelihood to confirm Vakkari's realization about the possibility to recognize new concepts added to the user's mental model by computing the number of search terms used [71, 72]. In order to verify his insight, it was decided to use the Term Criterion Labels from the Exploratory Search KiP model to check how a session evolved using the reformulation of its queries as a "roadmap". The idea was to map the query's reformulation from a dataset containing Exploratory Search Episodes, checking how they evolved from a generalization to a specialization pattern – such as generic (K1) to specific (K2), specific (K2) to specific not related (K3), generic (K1) to generic not related (K4), generic (K1) to specific not related (K3) – and vice-versa.

The researcher's understanding at this point was that the generalization/specialization and specialization/generalization perspectives could be used to infer implication to search quality. For example, relevance of the information retrieved, adequacy of results to search intention, user´s lack of knowledge and searching savviness compromising the meet of his/her intention.

For this experiment, a dataset was composed of selecting sessions and queries collected from transaction logs of an online teacher professional community used to validate the Exploratory Search KiP model's applicability to situations with less and in-depth level of context knowledge [67]. The community's name is LearnWeb-OER[10], so the composed dataset was named as the Learn Web Dataset[11]. The only modification done to the data was the changing of the original alphanumeric session ID, shown in Figure 16, to a more easily identified session name, e.g., LearnWeb01, LearnWeb02. The Learn Web dataset was comprised by 646 queries divided in 159 sessions.

[10] https://learnweb.l3s.uni-hannover.de/

[11] The dataset is available at http://doi.org/10.6084/m9.figshare.7637456

During the classification process, it was noted that the Term Criterion Labels was not enough to account for the entireness of query reformulation possibilities, generating a modification to the artifact of the first research cycle (Exploratory Search KiP model). For instance, the labels K1, K2, K3 and K4 could not properly differentiate the initial state of distinct search sessions presented by an Exploratory Search Episode. Neither could represent transitions more refined and that could not be accounted as specialization and generalization, such as the modification of certain words within a term without actually changing the query length. There was also trouble to represent situations like the ones displayed by the semantic related terms ‘spyware’ and ‘trojan horse’ and the return to the initial state query term during the same search session, the ‘lake como’ example presented at section 3.2.

To proper identify the types of query reformulation according to the needs reported at the previous paragraph, we propose the ESKiP Taxonomy of Query States. As a result of the collaboration between Colombian and Brazilian universities, the ESKiP Taxonomy is an adaptation of the existing classification of query reformulation originally created by Liu *et al.* [32], added by an initial state of a search – which can be a generic or specific term alike – and the return to initial state after several modifications. It also added the related reformulation type, represented by a change of term which although is different syntactically from the previous term it is not semantically. The taxonomy replaces the Term Criterion Labels at Exploratory Search KiP model’s Decision Diagram and is presented in Table 2.

Table 2: ESKiP Taxonomy of Query States.

Query State	Definition	Example
Initial State (IS)	Qi contains a set of terms representing the start of a search.	Lake Como
Return State (RS)	Qi contains at least one term and represents the start of a search or a previous search query; Qi+n contains exactly the same terms of Qi.	Lake Como -> Lake Como Vacation -> Lake Como
Generalization (GE)	Qi and Qi+1 contain at least one term in common; Qi+1 contains fewer terms than Qi.	Brazilian Flag Colors -> Brazilian Flag
Specialization (SC)	Qi and Qi+1 contain at least one term in common; Qi+1 contains more terms than Qi.	Brazilian Flag -> Brazilian Flag Colors
Repeat (RP)	Qi and Qi+1 may contain exactly the same terms, but the format of these terms may be different.	Brazilian Flag -> Flag Brazilian
Word Substitution (WS)	Qi and Qi+1 contain at least one term in common; Qi+1 might has the same length as Qi, but contains some terms that are not in Qi.	Brazilian Flag -> Colombian Flag
New (NW)	Qi and Qi+1 do not contain any common terms.	Brazilian Flag -> Dog Breeds
Related (RE)	Qi and Qi+1 do not contain common terms, but have similar or related meaning.	The Girl with the Dragon Tattoo Actress -> Rooney Mara images

Qi and Qi+1 have the same definition as depicted by Liu *et al.* [32]. Qi represents a current query and Qi+1 is the next following one. Qi+n represents a query used during the same search session than Qi but not necessarily the query immediately following. It is important to clarify that both Generalization (GE) and Specialization (SC) States are associated to reductions and increases regarding subject's domain not semantics. These fluctuations are based on query length and common words comparison.

The Related State (RE) introduced by ESKiP Taxonomy is also a response to a problem proposed by Liu *et al.* [32], referring to the identification of Query States when considering as features common term usage and the change of query length in the reformulated queries– an approach also used by Jansen, Booth and Spink [25]. The specific problem concerns the classification of the query reformulation type 'New'. Liu *et al.* [32] cited the possibility of the 'New' terms be related to the previous queries instead of being different ones, for example, by being composed of acronyms or synonyms.

A similar situation were also brought while analyzing queries that were generalizations and specializations that do not shared common terms with each other. Note that is the same specificity cited earlier and characterized by the use of syntactically different but semantically related terms. An example of this situation would be the queries 'While Loop' and 'Control Flow Statements'. Since it was opted not to define rules regarding specification or generalization from related terms, meaning not to consider a subdivision as Related Specialization or Related Generalization, the second query would be labeled as Related (RE).

Whether the queries from the previous example had a common term, such as 'Control Flow Statements' and 'For Loop Statements', they would characterize a Word Substitution (WS). It was also opted not to define a "Stemming State", meaning to reduce the words by their root, therefore whether the queries from the example were something like 'Control Flow Statements' and 'For Loop State' (root: Statement >> State), they would also be considered Related (RE).

Another example of the Related (RE) Query State situation, this time taken from one of the datasets used in the current research[12], is the following queries:

- q1 = ‘actress in girl with the dragon tattoo’ – labeled as New (NW) because the previous query was ‘paula deen’ and;
- q2 = ‘rooney mara images’ – labeled as Related (RE), because Rooney Mara is the "actress in girl with the dragon tattoo" movie.

It is important to stress that the Related (RE) Query State demands a different approach from the traditional common term / query length to be able to be identified as a reformulation type.

3.4. Method

To test the applicability of ESKiP Taxonomy of Query States as a tool to help search systems to identify query reformulation types, a classifier was implemented. Due to research restrictions, it was opted to avoid for the time being the attempt to identity the Related (RE) Query State. In future works, a worthwhile approach would be to train Machine Learning models to "learn" the human notion of similarity or relatedness, using semantic annotations, and apply this Artificial Intelligence to the identification task of classifying the particular Query State.

It was applied string similarity calculation to compare the classifier’s results to the original ones. String similarity is a measure that quantifies the similarity between two text

[12] The dataset will be properly present at section 3.5.

strings for approximate string matching or comparison [34]. To be able to execute the measurement, it is necessary to apply the mathematical concept of triangle inequality [41], which states “that for any triangle, the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side” [40], written as:

$$z \leq x + y$$

Triangle inequality is a theorem about distances and in Euclidean geometry can be written using vectors and vector lengths to represent the inequality of the triangle sides’ lengths x, y, z as:

$$||x + y|| \leq ||x|| + ||y||$$

The length z of the third side is replaced by the vector sum $x + y$ and when considering x and y real numbers, they can be viewed as vectors in ℝ1. In this case, the triangle inequality expresses a relationship between absolute values [40] and can be measured.

To apply the theorem to string similarity, it was used as method the Levenshtein distance, also referred as edit distance. The Levenshtein distance between two strings is calculate based on the minimum number of single-character edits, such as insertions, deletions or substitutions, required to change one word into the other [41]. Mathematically, it is represented as:

$$lev\ a, b(i,j) = \begin{cases} \max(i,j) \\ \min\begin{cases} lev\ a{,}b(i-1{,}j)+1 \\ lev\ a{,}b(i{,}j-1)+1 \\ lev\ a{,}b(i-1{,}j-1)+1(ai{\neq}bj) \end{cases} \end{cases} \quad if min(i,j) = 0,$$

$$otherwise.$$

$lev\ a, b(i, j)$ is the distance between the first i characters of a and the first j character of b.

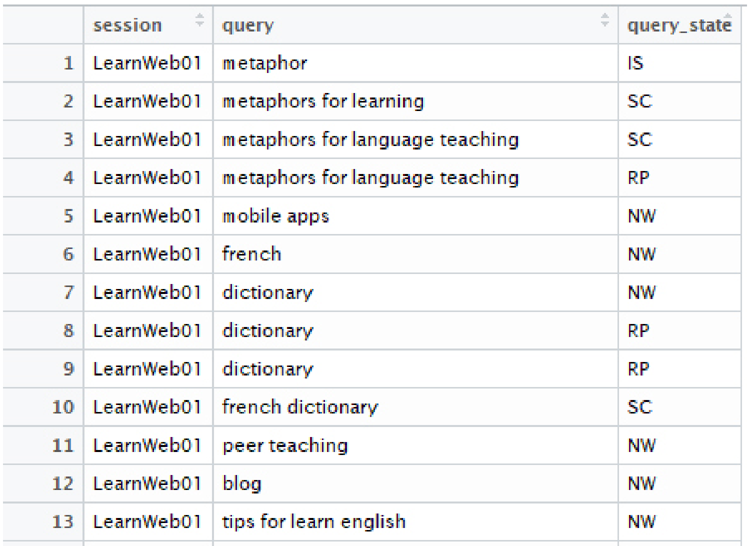
$1(ai \neq bj)$ is the characteristic function[13] equal to 0 when $ai = bj$ and equal to 1 otherwise.

Any edit required adds a point to the calculation and composes the distance between the strings. The first element in the minimum corresponds to deletion, the second refers to insertion and the third defines concordant and discordant matches, which determines the string similarity calculation to positive (1) and negative (0).

3.5. Evaluation Setup

The proposed Query States were manually identified at the Learn Web dataset to examine the usage of the Taxonomy's query reformulation types (Figure 22). As Liu *et al.* cited in [32], it is important to understand "how people reformulate their queries in different tasks or under different situations", a behavior defined at this thesis as User Search Strategies, because such knowledge could be used to help search systems to provide better query suggestions and improve search results.

[13] Indicates membership of an element in a subset A of X, having the value 1 for all elements of A and the value 0 for all elements of X not in A.



	session	query	query_state
1	LearnWeb01	metaphor	IS
2	LearnWeb01	metaphors for learning	SC
3	LearnWeb01	metaphors for language teaching	SC
4	LearnWeb01	metaphors for language teaching	RP
5	LearnWeb01	mobile apps	NW
6	LearnWeb01	french	NW
7	LearnWeb01	dictionary	NW
8	LearnWeb01	dictionary	RP
9	LearnWeb01	dictionary	RP
10	LearnWeb01	french dictionary	SC
11	LearnWeb01	peer teaching	NW
12	LearnWeb01	blog	NW
13	LearnWeb01	tips for learn english	NW

Figure 22: A sample from Learn Web dataset.

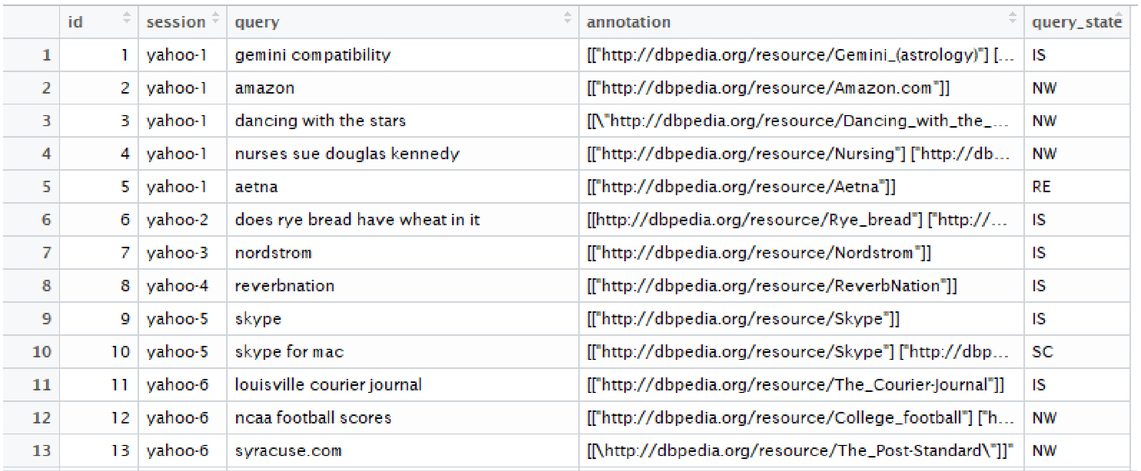
Query reformulations are direct correlated to the amount of information needed to complete a search task [32]. As more query reformulations were issued during searching, the cognitive effort spent in exploring in more detail the information available increases. Liu *et al.* [32] determined a percentage of 47.19% in effective reformulation, meaning that more than half of user's attempts to refine the search results are non-effective. The amount of non-effective query reformulation, in this researcher view, justifies attempts to improve the identification of Query States and their transitions. It also led to the decision of capturing the frequency of each query reformulation type, not to infer implication to search quality, but to plan how to successfully identify and classify Query States. Table 3 shows the frequency of each Query State at the Learn Web dataset.

Table 3: Query States identified at Learn Web dataset.

Query State	Frequency	Overall Frequency
Initial State (IS)	159	24.61%
Return State (RS)	8	1.24%
Generalization (GE)	17	2.63%
Specialization (SC)	40	6.19%
Repeat (RP)	278	43.03%
Word Substitution (WS)	17	2.63%
New (NW)	83	12.85%
Related (RE)	44	6.81%

To validate the Taxonomy and check the consistency of query reformulation types, it was necessary to verify the appearance of all proposed Query States in a different set of Exploratory Search Episodes. To this purpose, Yahoo![14] provided a dataset comprising of queries taken from Yahoo US Web Search and from the TREC Session track (2010-2013) with 2633 queries in 845 sessions, 7482 spans, and 5964 links to Wikipedia articles (Figure 23). Again, the Query States were manually identified and the eight proposed query reformulation types had their frequency mapped, as seen in Table 4.

[14] Yahoo! Webscope dataset ydata-search-query-log-to-entities-v1_0
[http://labs.yahoo.com/Academic_Relations]



	id	session	query	annotation	query_state
1	1	yahoo-1	gemini compatibility	[["http://dbpedia.org/resource/Gemini_(astrology)"] [...	IS
2	2	yahoo-1	amazon	[["http://dbpedia.org/resource/Amazon.com"]]	NW
3	3	yahoo-1	dancing with the stars	[[\"http://dbpedia.org/resource/Dancing_with_the_...	NW
4	4	yahoo-1	nurses sue douglas kennedy	[["http://dbpedia.org/resource/Nursing"] ["http://db...	NW
5	5	yahoo-1	aetna	[["http://dbpedia.org/resource/Aetna"]]	RE
6	6	yahoo-2	does rye bread have wheat in it	[[http://dbpedia.org/resource/Rye_bread"] ["http://...	IS
7	7	yahoo-3	nordstrom	[["http://dbpedia.org/resource/Nordstrom"]]	IS
8	8	yahoo-4	reverbnation	[["http://dbpedia.org/resource/ReverbNation"]]	IS
9	9	yahoo-5	skype	[["http://dbpedia.org/resource/Skype"]]	IS
10	10	yahoo-5	skype for mac	[["http://dbpedia.org/resource/Skype"] ["http://dbp...	SC
11	11	yahoo-6	louisville courier journal	[["http://dbpedia.org/resource/The_Courier-Journal"]]	IS
12	12	yahoo-6	ncaa football scores	[["http://dbpedia.org/resource/College_football"] ["h...	NW
13	13	yahoo-6	syracuse.com	[[\http://dbpedia.org/resource/The_Post-Standard\"]]"	NW

Figure 23: A sample from Yahoo! Dataset.

Table 4: Query States identified at Yahoo! Dataset.

Query State	Frequency	Overall Frequency
Initial State (IS)	845	32.09%
Return State (RS)	6	0.23%
Generalization (GE)	91	3.46%
Specialization (SC)	324	12.31%
Repeat (RP)	79	3.00%
Word Substitution (WS)	529	20.09%
New (NW)	551	20.93%
Related (RE)	208	7.90%

An observable result is the appearance of each Query State in both datasets. The Initial State (IS) marks the start of a session and, as expected, accounts for the number of each session within the datasets. The overall frequency of query reformulation types indicates different Search Profiles in the datasets. The overall frequency was calculated by using the number of queries in each dataset to determine the percentage of each Query State by the frequency of its appearance, e.g., in Learn Web dataset SC appeared 40 times from 646 queries, hence an overall frequency of 6.19%.

Rha, Shi and Belkin [16] determined specialization as the most frequent information reason to reformulate, but in Yahoo! dataset it was responsible for 12.31% of query reformulations while at Learn Web dataset accounted for only 6.19%. The most frequent Query State present at Learn Web dataset is Repeat (RP), a choice made in 43.03% of occasions. It means that users preferred repeat times in a row the same query. In the meantime, at Yahoo! dataset the most frequent reason to reformulate was to start a New (NW) query in the same session (20.93%)[15].

To implement the classifier, three groups of functions were composed using pairs of queries as input. A first function to separate each word used to compose the terms presented in the queries and counting them. A second function to compare the pair of queries, identify duplicate words, i.e., common words in both queries, and return the number of duplicate words as a response. A third one uses the two previous functions and the description of the Taxonomy's Query States to return the classification as output. The pseudocode to explain the functions' logic is a mixture of English and Python and is shown below, as well as a proper explanation to each feature.

[15] Note that Initial State (IS) could not be considered a reason to reformulate because it is a mandatory Query State.

Table 5: Word Count Function.

Word Count Function: to split the query by words and count them

```
procedure wordCounts (query string):
   to store counting use counts ← create a dict Hash Table
   words ← split the query string

   for word in words:
      if word in counts:
         counts[word] += 1
      else:
         counts[word] = 1

   return (counts)
end procedure
```

The Python language has an efficient key/value hash table structure called "dict", which contents can be written as a series of key:value pairs. The function (Table 5) uses the query string as input and starts such structure to store the words and their counting (e.g. dict = {key1:value1, key2:value2}), then split the query string in words. For instance, considering a pair as [q1,q2] with q1 = 'start query' and q2 = 'following query' as inputs, the function's output would be:

```
Out [q1]: {'query': 1, 'start': 1}

Out [q2]: {'following': 1, 'query': 1}
```

Table 6: Word Duplicate Function.

Word Duplicate Function: to compare a pair of queries and count their common words

```
procedure wordDuplicates(query string 1, query string 2):
   start count in 0
   words ← split the query string 1

   for word in query string 2:
      if word in words:
         count += 1

   return (count)
end procedure
```

The second function (Table 6) considers as input the query string pair to be compared, then split each query string and compares the second to the first one. If there is a common word, it returns the number of common words. If there is none, the number returned is zero. Considering the last example presented, the function output would be Out [q1,q2]: 1 – as the only common word in them is ‘query’.

Table 7: Query State Classifier Function.

Query State Classifier Function: to classify a query based on ESKiP Taxonomy of Query States

```
procedure QueryStateClassifier(query string):

   while query string in Qi:
      Qi ← the set containing initial state query strings
       for query string ∈ Qi:
          classify as “IS”

   while query string in Qm:
      Qm ← the set containing previous state query strings
       for query string ∈ Qm:
         if length of Qm < length of Qi and wordDuplicates > 0:
            classify as "GE"
         elif length of Qm > length of Qi and wordDuplicates > 0:
            classify as "SC"
         elif wordCounts to Qm == wordCounts to Qi:
            classify as "RP"
         elif wordCounts to Qm != wordCounts to Qi and wordDuplicates > 0:
            classify as "WS"
         elif wordCounts to Qm != wordCounts to Qi and wordDuplicates > 0 and
              length of Qm < length of Qi:
            classify as "WS"
         elif wordCounts to Qm != wordCounts to Qi and wordDuplicates > 0 and
             length of Qm > length of Qi:
            classify as "WS"
         elif wordCounts to Qm != wordCounts to Qi and wordDuplicates == 0:
            classify as "NW"
         elif wordCounts to Qm != wordCounts to Qi and wordDuplicates == 0 and
              length of Qm < length of Qi:
            classify as "NW"
         elif wordCounts to Qm != wordCounts to Qi and wordDuplicates == 0 and
              length of Qm > length of Qi:
            classify as "NW"
         else:
            classify as "RE"
```

```
  while query string in Qn:
    Qn ← the set containing current state query strings
    for query string ∈ Qn:
      if wordCounts to Qn == wordCounts to Qi and wordDuplicates > 0:
        classify as "RS"
      elif length of Qn < length of Qm and wordDuplicates > 0:
        classify as "GE"
      elif length of Qn > length of Qm and wordDuplicates > 0:
        classify as "SC"
      elif wordCounts to Qn == wordCounts to Qm:
        classify as "RP"
      elif wordCounts to Qn != wordCounts to Qm and wordDuplicates > 0:
        classify as "WS"
      elif wordCounts to Qn != wordCounts to Qm and wordDuplicates > 0 and
          length of Qn < length of Qm:
        classify as "WS"
      elif wordCounts to Qn != wordCounts to Qm and wordDuplicates > 0 and
          length of Qn > length of Qm:
        classify as "WS"
      elif wordCounts to Qn != wordCounts to Qm and wordDuplicates == 0:
        classify as "NW"
      elif wordCounts to Qn != wordCounts to Qm and wordDuplicates == 0 and
          length of Qn < length of Qm:
        classify as "NW"
      elif wordCounts to Qn != wordCounts to Qm and wordDuplicates == 0 and
          length of Qn > length of Qm:
        classify as "NW"
      else:
        classify as "RE"

  return (Qi, Qm, Qn)
end procedure
```

The classification function (Table 7) considers as input a query string and compares it to the previous one, assuming the Markov property presented in probability theory. In the event of no previous query string, the function compares the query to itself to declare it as the search's initial state. It works within three types of states: Qi is the initial state; Qm is a hybrid state, acquiring a role of following state to Qi and a role of previous state to Qn; and finally Qn the current state. When Qi is determined, it is assigned the label IS for Initial State and the function proceeds to the second query string. Qm is then determined by the following instructions:

1) Check the length of Qm and Qi and use the function wordDuplicates to compare the pair of queries:

- If Qm has less words than Qi and has at least one word in common, assign the label GE, for Generalization.
- If Qm has more words than Qi and has at least one word in common, assign the label SC, for Specialization.

2) Use the function wordCounts to Qm and Qi and compare its result. If the result is the same for both, assign the label RP, for Repeat.

3) Use the function wordCounts to Qm and Qi, compare its result and then use the function wordDuplicates to compare the pair of queries. Also, check the length of Qm and Qi:

- If the wordCounts to Qm is different from the wordCounts of Qi and they have at least one word in common, assign the label WS, for Word Substitution.
  - ✓ If the previous analysis was accompanied by Qm having less words than Qi or Qm having more words than Qi, also assign the label WS.
- If the wordCounts to Qm is different from the wordCounts of Qi and they have no words in common, assign the label NW, for New.
  - ✓ If the previous analysis was accompanied by Qm having less words than Qi or Qm having more words than Qi, also assign the label NW.

4) Any result different from the above, assign the label RE, for Related.

Qn is determined by all the four previous instructions – the only difference being the changing of the comparison pair to Qn and Qm instead of Qm and Qi – and by the addition of one more direction before the instruction number four:

- Use the function wordCounts to Qn and Qi, compare its result and then use the function wordDuplicates to compare the pair of queries. If Qn wordCounts is the same as Qi wordCounts and they have the same common words, assign the label RS, for Return State.

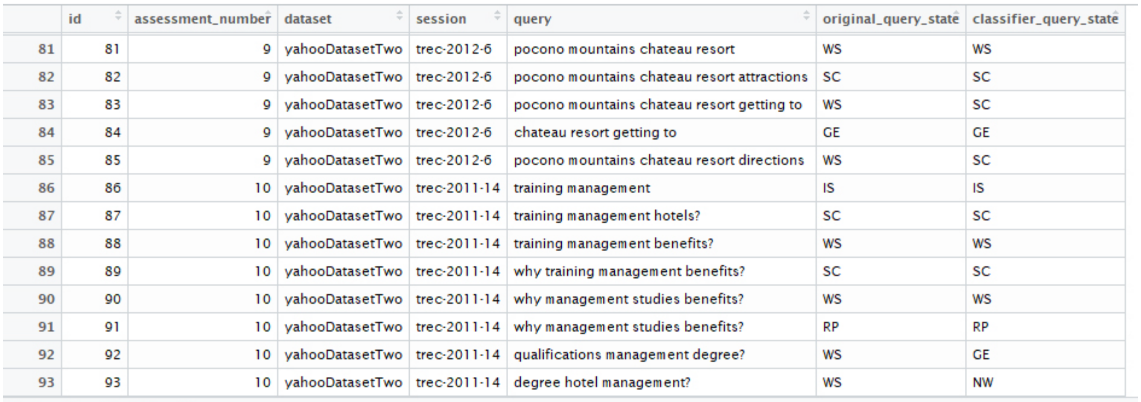
It is necessary to acknowledge that the inclusion of the forth instruction had only the slightest chance of succeeding and accounted for a “long shot” attempt to identify the Related (RE) Query State without adding semantic power to the classifier.

3.6. Results

The classifier was applied to ten search sessions retrieved from the Yahoo! dataset. The sessions was chosen based on three criterion defined by the researcher:

1) It had to display an indication of exploratory search pattern by having at least five queries;
2) It had to display an indication of a search strategy performed by having at least two Query States transitions;
3) It could not be consecutive sessions to avoid the possibility of had being performed by the same user.

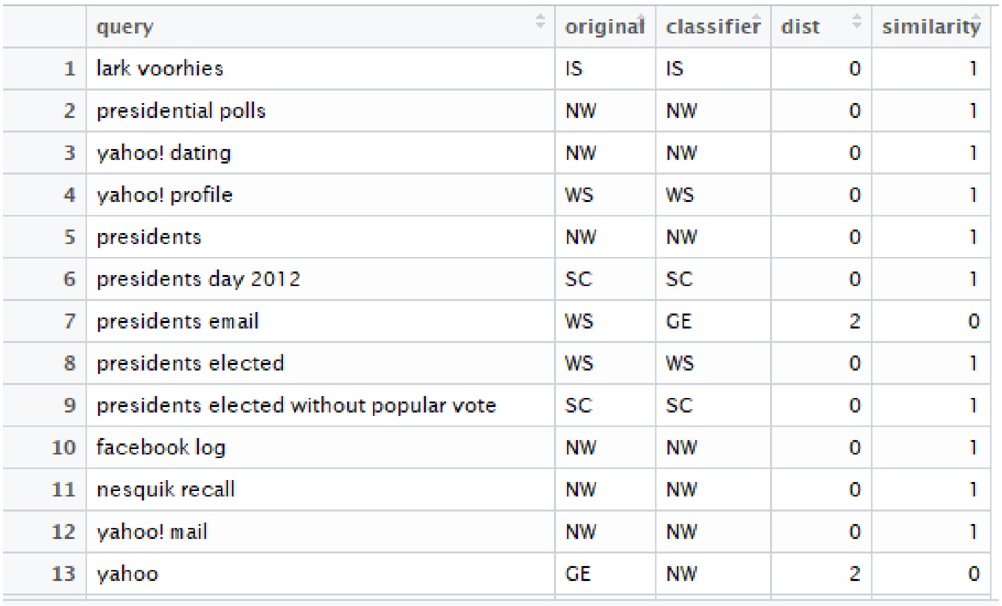
Figure 24 shows a sample of the original manually identified Query States and the classifier’s results.



	id	assessment_number	dataset	session	query	original_query_state	classifier_query_state
81	81	9	yahooDatasetTwo	trec-2012-6	pocono mountains chateau resort	WS	WS
82	82	9	yahooDatasetTwo	trec-2012-6	pocono mountains chateau resort attractions	SC	SC
83	83	9	yahooDatasetTwo	trec-2012-6	pocono mountains chateau resort getting to	WS	SC
84	84	9	yahooDatasetTwo	trec-2012-6	chateau resort getting to	GE	GE
85	85	9	yahooDatasetTwo	trec-2012-6	pocono mountains chateau resort directions	WS	SC
86	86	10	yahooDatasetTwo	trec-2011-14	training management	IS	IS
87	87	10	yahooDatasetTwo	trec-2011-14	training management hotels?	SC	SC
88	88	10	yahooDatasetTwo	trec-2011-14	training management benefits?	WS	WS
89	89	10	yahooDatasetTwo	trec-2011-14	why training management benefits?	SC	SC
90	90	10	yahooDatasetTwo	trec-2011-14	why management studies benefits?	WS	WS
91	91	10	yahooDatasetTwo	trec-2011-14	why management studies benefits?	RP	RP
92	92	10	yahooDatasetTwo	trec-2011-14	qualifications management degree?	WS	GE
93	93	10	yahooDatasetTwo	trec-2011-14	degree hotel management?	WS	NW

Figure 24: A sample from the classifier results.

Figure 25 shows a sample of the string similarity calculation comparing the original results to the classifier's.



	query	original	classifier	dist	similarity
1	lark voorhies	IS	IS	0	1
2	presidential polls	NW	NW	0	1
3	yahoo! dating	NW	NW	0	1
4	yahoo! profile	WS	WS	0	1
5	presidents	NW	NW	0	1
6	presidents day 2012	SC	SC	0	1
7	presidents email	WS	GE	2	0
8	presidents elected	WS	WS	0	1
9	presidents elected without popular vote	SC	SC	0	1
10	facebook log	NW	NW	0	1
11	nesquik recall	NW	NW	0	1
12	yahoo! mail	NW	NW	0	1
13	yahoo	GE	NW	2	0

Figure 25: A sample from the string similarity calculation.

Taking as an example the row 2 presented at Figure 25, the Levenshtein distance between the original and the classifier's classification is 0 since there is no need to perform single-character edits because both classifications are New (NW). Thus, assigned a positive (1) similarity. Row 7 shows a different situation, the original classification was set as Word Substitution (WS) while the classifier determined it as a Generalization (GE).

The Levenshtein distance between both is 2, once it was necessary two edits to change one into the other and there is no alternative to do it with fewer than this amount of edits. The procedure was:

1. WS -> GS (substitution of “W” for “G”)
2. GS -> GE (substitution of “S” for “E”)

Thus, row 7 was assign a negative (0) similarity. After applying the proceeding to every query classified from the ten chosen sessions, the researcher were able to determine the String Similarity Rate by computing the percentage of concordant and discordant matches, as shown in Table 8:

Table 8: String Similarity Rate.

Similarity Type	Rate
Positive (1)	0.7
Negative (0)	0.3

3.7. Conclusion

The negative (0) rate of 0.3 was due to some dissonant results from both classification, as can be noted at Table 9, which compares their Query States Frequency.

Table 9: Results comparison: Query States Frequency.

Query State	Original Query State Frequency	Classifier Query State Frequency
Initial State (IS)	10	10
Return State (RS)	N/A	N/A
Generalization (GE)	3	8
Specialization (SC)	16	27
Repeat (RP)	3	1
Word Substitution (WS)	35	17
New (NW)	21	30
Related (RE)	5	N/A

The classifier identified more Query States as Generalization (GE), Specialization (SC) and New (NW) while identified less as Repeat (RP) and Word Substitution (WS). The Return State (RS) was not present at the analyzed sample and, as expected, the Related (RE) state is not identifiable without the addition of semantics to the classifier.

The dissonance was caused by two weaknesses from this classifier's version:

1) A difficulty in identifying Word Substitution (WS) state when the length, i.e., the number of words, of the compared two queries are different. It identifies as Generalization (GE) or Specialization (SC), because the queries have words in common. It accounts for the increase of the amount of GE and SC states and the decrease of the WS state identified.
2) Another difficulty perceived is in differentiating the same word written in upper and lower case – e.g. "Yahoo" and "yahoo" – with some symbol, e.g., "yahoo" and "yahoo!" or plural, e.g., "papa john" and "papa johns". As the classifier identifies them as different words, it interferes the correct

classification of the Query State in consecutive queries that contain situations such as described.

To improve the positive (1) String Similarity Rate of 0.7 and strengthen the classifier's ability to overcome the weaknesses described above, the inclusion of a new function to tune the Generalization (GE) and Specialization (SC) cases is necessary. It could set a rule to define the cited Query States based on a string similarity measure, such as the one used to evaluate the classifier, and refine its ability to differentiate them from Word Substitution (WS).

Another necessary update is the inclusion of a Machine Learning model to the classifier to provide it with a capability to “understand” the human notion of similarity and uses it to identify the Related (RE) state. A possibility of such implementation is the application of the Semantic Connective Score, developed by Nunes *et al.* [45], that uses semantic annotations to define a score between entities based on the description of their relations. The SCS could be calculate considering two different max path lengths, e.g., max path lengths of 1 and 2, and be employed as a feature to a Machine Learning model using training methods such as Logistic Regression, Naïve Bayes or Decision Tree. The model could also be used to subdue the discernment problem concerning the same word written with the differences described earlier as well as the implementation of functions to reduce words to its radical and to transform the string to lowercase.

4. CONCLUSION AND FUTURE RESEARCH

4.1. Findings on Exploratory Search KiP model and ESKiP Taxonomy of Query States

Investigating Knowledge-intensive activities from Exploratory Search KiP model ("Search Term Selection", "Query Formulation", "Results Check" and "Information Extraction") helps visualize the behaviors involved in an exploratory search process and in describing how they relate to knowledge acquisition, sharing, storing and reusing. It is possible to distinguish three main behaviors involved in the mentioned activities.

The first, observable at the beginning when the user becomes familiar with the subject, is important to reduce the uncertainty involved in the activity. This subject familiarization makes it possible to select a search term. The second behavior involves the ability to control the search process itself and it increases as it moves from the first KiP activity to the second. Queries formulation and reformulation, formalized by search string definition, indicates there is a new knowledge stage regarding the topic. It is not possible to define a string if there is no conceptual understanding whatsoever about the searched subject. It is precisely this understanding that indicates an ability to control the process and prepares the user for the KiP activities that follows.

The third behavior encompasses the remainder of the Exploratory Search KiP model and is characterized by an ability to assess the retrieved information relevance. For instance, it is observable in the contingent event associated to Results Check activity –

information visualization – that aids in defining the main criteria to help the user choose what information sources will be used. This evaluation capability influences what information the user selects, extracts and stores and might contribute to indicate if learning occurred during the process.

In conclusion, the Exploratory Search KiP model can be applied to mind map search patterns and learning processes useful to two perspectives: (a) computer developers and (b) teachers and educators. As to the first perspective, understanding search processes and patterns is important to fields connected to human-machine interaction, information retrieval, information science and semantic-based search. Understanding how user intention evolves during a search process is pivotal to assign meanings to concepts and evaluate the quality of the retrieved information based on the proper understanding of the user's initial search intention.

As to the second perspective, the model supports expertise exchange and can be used in studies in which expert domain has a pivotal role in the process. As it enabled the present study to visualize the teachers' search pattern and map their decision-making process, Exploratory Search KiP model might be used as an aid to understand learning processes related to educational technology, online education and self-learning based on the internet. The model could also support the development of a new generation of intelligent tutoring systems, grounded on the understanding of learners' decision-making while interacting with the system. It can also be used at the identification of Query States – as in the second research cycle – to provide more accurate results rooted on the understanding of the patterns employed to reformulate the terms used to form the queries.

From the application of the Exploratory Search KiP Model to log analysis, this research were able to identify four important characteristics of users' decision-making processes while searching online:

(1) As seen in searching actions that prompted downloads without content checking, the decision criteria about which data are worth extracting is internalized and tacit as users get more acquainted with the subject by analyzing the information retrieved in each search.

(2) As seen in the Term Criterion Label Sequence shown as example, there can be a refinement from general term (K1) to a specific one (K2), then a change of subject (K4) and the repetition of this pattern until users reach a decision to adopt a more specific subject not related to the general term previously used (K3). This means that users adopt different search strategies to draw inferences and devise shortcuts in order to improve their understanding.

(3) The Term Criterion Label Sequence also helps to understand the formation and refinement of users' mental image about the subject and its impact on their decision process. The term refinement from general to specific, even a specification not related to the previous general term, indicates that as users acquire deeper awareness, possibly from analyzing the search results, their search intention is modified, as shown by a persistent change of subject.

(4) As seen in the sample of an Exploratory Search KiP modeled for one of the sessions (Figure 15), the purpose that drives the search turns into real search actions influenced by the selection of terms and their ensuing refinement.

As for the ESKiP Taxonomy of Query States it is an update to Liu *et al.'s* [32] five categories taxonomy of query reformulation – Generalization, Specialization, Repeat, Word Substitution and New – with the inclusion of three new Query States – Initial State, Return State and Related. The introduction of the Taxonomy provides more refinement possibilities to the identification of Query States and their transitions. It opens more opportunities in the field of knowledge representation and reasoning in helping to represent the structure and behavior of queries reformulation in search systems.

By analyzing the overall frequency of Query States present in both datasets used in this research (Tables 3 and 4), it is possible to draw some conclusion regarding user behavior. Since the researcher can only assume users intentions – the datasets' data did not provided such information – based on the data's origins it seems plausible that search intention did play a role in define users' search strategies. Learn Web dataset came from a collaborative search and sharing system that links different online services and provides advanced support to students and teachers organize and share the resources they find online. Its searches demonstrate more exploratory pattern, in which the information available needs to be explored in more detail, than Yahoo! dataset's. The search engine logs supplied by Yahoo! reflect a less exploratory pattern – and maybe it explains the reason behind the users' behavior of changing search subject so often during the same session, as commented at section 3.5.

Word Substitution (WS) accounting for 20.09% of queries reformulation at Yahoo! and 2.63% at Learn Web corroborates to the inference derived from the data's origin. A User Search Strategy based on WS indicates a tendency of refine search results by finding the words that could prompt the best results. As is the case, for instance, of q1 = 'yahoo mail india' and q2 = 'yahoo mail plus' – taken from Yahoo! dataset. In Learn Web dataset there is a preference to adding more terms to the query (6.19%) or finding related terms (6.81%), since the Specialization (SC) and Related (RE) Query States are the most frequent choices of search refinement without entirely changing the subject and indicates a greater awareness about searched topics than the WS strategy.

In the Learn Web dataset is observed a predilection for the term repetition (RP) in consecutive queries (43.03%). It indicates that users may want to revisit information or even find or check for new information on topics they have previously explored. This information might be employed, e.g., to determined context differences based on User Search Strategy and assort a learning purpose to the searches performed there.

The identification of terms' sequence by the classification of query reformulation types and their labeling to Query States helps to visualize the search strategies users adopt. It can be used to define Search Profiles based on users' search shrewdness and apply the profiles to design search interfaces to better support learning with respect to the search process, noted by Rieh *et al.* [59] as one of the main topics in Searching as Learning research agenda. Terms suggestions to compose queries, tips to help the user improve its search abilities and search paths' recommendation to users are examples of potential applications.

A possible impact in artificial intelligence applied to Web search engines is in formalizing semantic networks[16] that could be used to better represent relations between concepts based on the query terms. As an expected outcome, it would provide more accurate results grounded on the semantic understanding of the terms used. A Natural Language Processing implementation based on ESKiP Taxonomy could improve semantic matching in Web searching, for example, by identifying the Related State (RE) of two consecutive queries.

[16] It is a knowledge base, modeled as a directed or undirected graph, which represents semantic relations between concepts in a network.

4.2. Limitations

Some inferences derived from the first research cycle are empirical. The tagging resource activity at the presented log analysis is a good example of this empiricism. It can be interpreted as a measure of user understanding or awareness about the subject searched and based on Search as Learning concepts could indicate the occurrence of learning during the user's exploration of the retrieved results. Its frequency shows that users that perform more search sessions tend to use tagging more often, as shown in Figure 17. Could this observation indicate learning happening through interaction with contents during each search, i.e., texts read, diagrams analyzed, and video watched? All the described inferences need more evidence regarding their likelihood.

Exploratory Search KiP model would benefit from its application in more case studies with more routinely search situations. For example, a question worth made is if it would map adequately searches with less or no exploratory pattern. Based on the Yahoo! dataset it seems to indicate a positive answer, but it needs further research on this matter.

The usage of the ESKiP Taxonomy in the development of a classifier to identify query reformulation is a demonstration of its applicability to real-case scenarios, with a possible impact on information retrieval improvement. Nevertheless, the lack of a semantic application to identify Related States (RE) is a major restriction. To be able to successfully influence on Web search engine's ability to recognize semantic networks it is pivotal that the Related State (RE) identification in query reformulation be viably implemented.

User Search Strategies appearing in both datasets indicate that Search Profiles may be connected to the situations that influence user's behavior and determine how the search activities are performed, e.g., lack of enough information to structure a proper query, user's search expertise, and new information added changing the search problem definition. However, more researches are necessary to provide more evidence on the subject.

4.3. Future Research

A future development foreseen concerning the impact on information retrieval – and information seeking as a whole – is the association of the ESKiP Taxonomy to entity-centric ranking in order to improve Web search results. Entity-centric data management is an area that has received increasing attention as a research field and encompasses a number of disciplines such as Databases, Information Retrieval, and the Semantic Web. Entity ranking has played a key role in information retrieval and used for such tasks as expert finding, where the goal is to find people who have expert knowledge about a particular topic. In recent years, works have looked at how to search for multiple entity types [68]. Standard information retrieval methods based on inverted indices are associated to structured search based on graphs to connect entities and to improve search effectiveness [68].

In Natural Language Processing, entity ranking is used to builds up clusters of mentions relying on partially formed clusters produced to make decisions regarding relationships. Then these clusters of mentions are merged if the ranking model predicts they are representing the same entity. In this way, the approach can successfully reject linking [Hillary Clinton] with [Clinton, he], for example, because of the low score between the pair [Hillary Clinton, he], as mentioned by Clark and Manning in [8].

Another approach, applicable to Web search engines, ranks entity-centric collections in order of relevance to a query, and then identify a set of collections that are likely to contain most relevant entities, known as Entity-Centric Collection Ranking (EC), in which the central broker, according to their probability of relevance, ranks entities. The top relevant entities contribute to the collection's query-likelihood score [1].

Entity-centric ranking is also associated to knowledge bases modeled as graphs to use their representations of entities, along with their related types, to rank the types assigned to entities from the hierarchy created by the graph. In an early work by Tonon *et al.* [68], this method was used to select the right granularity of types from the background type hierarchy. Usually, search effectiveness is achieved by combining approaches, especially those blending various ranked lists. The number of entities used for the graph-based search step influences search effectiveness, as seen in [6], in which Bron *et al.* employed a linear combination of the normalized similarity scores of the text for this purpose.

Semantic Connective Score (SCS) and Co-occurrence-Based Measure (CBM) are two other approaches that could be used in association to ESKiP Taxonomy of Query States to build information systems capable not only to retrieve relevant information but also to understand a Web search pattern and adjust the results accordingly. Contributing to the user's learning process as a result. Semantic Connective Score (SCS) is an index to estimate relatedness of actors in RDF graphs [3, 45] and Co-occurrence-Based Measure (CBM), is a measure that relies on an approximation of the number of Web pages that contain the labels [46] to estimate a score to a pair of entities.

Query States can be considered to identify query reformulation as it intersects research agendas regarding not only Information Retrieval and Information Seeking, but also Searching as Learning. With increasing costs of education systems and the shift from an industrial society to a knowledge-based society posing additional challenges in such arenas as ethics, economics, and working environment [65] it is not only necessary but also of vital importance to associate informal and self-learning to formal learning.

Therefore, societies around the globe could successfully prepare their citizens to operate in a new social context, in which cognitive skills to develop knowledge and intellectual property are highly considered and valued. In this regard, the internet appears as an essential tool because it contains a great part of the knowledge developed by humankind. Web search engines, in such context, gain importance and require to be improved either by the addition of semantic understanding or by the betterment of their ranking algorithms to be able to function as knowledge scaffoldings.

REFERENCES

[1] BALOG, KRISZTIAN; NEUMAYER, ROBERT; NØRVAG, KJETIL. “Collection Ranking and Selection for Federated Entity Search”. L. Calderon-Benavides et al. (Eds.): SPIRE 2012, LNCS 7608, pp. 73–85, 2012.

[2] BATES, M. J. “The design of browsing and berrypicking techniques for the online search interface”. *Online Review*, 13(5), pp. 407–424. 1989.

[3] BINGI, R., KHAZANCHI, D., & YADAV, S. B. “A framework for the comparative analysis and evaluation of knowledge representation schemes”. *Information Processing and Management*, 31(2), 233-247. 1995.

[4] BLOOM, B. S.; ENGELHART, M. D.; FURST, E. J.; HILL, W. H.; KRATHWOHL, D. R. “Taxonomy of educational objectives: The classification of educational goals”. *Handbook I: Cognitive domain*. New York: David McKay Company. 1956.

[5] BORTOLUZZI, M.; MARENZI, I. “Web searches for learning: how language teachers search for online resources”. *Lingue e Linguaggi*, 23, pp. 21-36. 2017.

[6] BRON, MARC; BALOG, KRISZTIAN; DE RIJKE, MAARTEN. “Example based entity search in the web of data”. In: *ECIR*, pp. 392–403. 2013.

[7] CHOO, C., DETLOR, B., &TURNBULL, D. “A behavioral model of information seeking on the Web: Preliminary results of a study of how managers and it specialists use the Web”. In C.M. Preston (Ed.), *Proceedings of the 61st Annual Meeting of the American Society for Information Science (ASIS)* (pp. 290–302). Medford, NJ: Information Today. 1998.

[8] CLARK, KEVIN; MANNING, CHRISTOPHER D. “Entity-Centric Coreference Resolution with Model Stacking.” *Assoc. Comput. Linguist*. 2015.

[9] DAVENPORT, TH., *Thinking for a living: how to get better performance and results from knowledge workers*. Harvard Business Review Press, Boston. 2005.

[10] DAVENPORT TH, JARVENPAA SL, BEERSMC. “Improving knowledge work processes”. *Sloan Manag Rev* 37(4):53–65. 1996.

[11] DAVIS EDUCATIONAL FOUNDATION. *An Inquiry into the Rising Cost of Higher Education: Summary of Responses from Seventy College and University Presidents.* In: Report Davis Educational Foundation. November 2012. Accessed in November 2017.
http://www.davisfoundations.org/site/documents/AnInquiryintotheRisingCostofHigherEducation_003.pdf

[12] DI CICCIO, CLAUDIO & MARRELLA, ANDREA & RUSSO, ALESSANDRO. "Knowledge-Intensive Processes: Characteristics, Requirements and Analysis of Contemporary Approaches". *Journal on Data Semantics*. 4. 29-57. 2015. 10.1007/s13740-014-0038-4.

[13] DI CICCIO C, MARRELLA A, RUSSO A. "Knowledge-intensive processes: an overview of contemporary approaches". In: *KiBP'12*. 2012.

[14] ELLIS, DAVID; COX, DEBORAH, HALL, KATHERINE. "A comparison of the information seeking patterns of researchers in the physical and social sciences". *Journal of Documentation*. 49 (4): 356–369. 1993.

[15] ERDEM SARIGIL, ISMAIL SENGOR ALTINGOVDE, ROI BLANCO, et al. "Characterizing, predicting, and handling web search queries that match very few or no results". *Journal of the Association for Information Science and Technology*. Volume 69, Issue2. Pages 256-270. 2018.

[16] EY RHA, W SHI, NJ BELKIN. "An exploration of reasons for query reformulations". *Proceedings of the Association for Information Science and Technology*, Wiley Online Library, Volume54, Issue1. Pages 337-346. 2017.

[17] FRANÇA, JULIANA BAPTISTA DOS SANTOS ; NETTO, JOANNE MANHÃES; DO E. S. CARVALHO, JULIANA ; et al.; . "KIPO: the knowledge-intensive process ontology". *Software & Systems Modeling,* v. 14, p. 1127-1157, 2015.

[18] GAUCH, S., & SMITH, J. "An expert system for automatic query reformulation". *Journal of the American Society for Information Science*, 44(3), 124–136. 1993.

[19] GIRARD, JOHN P; GIRARD, JOANN L. "Defining knowledge management: Toward an applied compendium". *Online Journal of Applied Knowledge Management.*

3 (1): 14. 2015.

[20] GOLDKUHL, G. “Anchoring Scientific Abstractions— Ontological and Linguistic Determination Following Socio-Instrumental Pragmatism,” *European Conference on Research Methods in Business and Management,* Reading, UK, pp. 29-30. 2002.

[21] GREGOR, S., HEVNER, A.: “Positioning and presenting design science research for maximum impact”. *MIS Q*. 37, 337–355. 2013.

[22] HEVNER, A.: “A three-cycle view of design science research”. *Scand. J. Inf. Syst.* 19(2), 87–92. 2007.

[23] HOWARD, P.N., & MASSANARI, A. “Learning to Search and Searching to Learn: Income, Education, and Experience Online”. *J. Computer-Mediated Communication,* 12, 846-865. 2007.

[24] INGWERSEN, P. “Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory”. *Journal of Documentation*, 52(1), 3–50. 1996.

[25] JANSEN, B.J., BOOTH, D.L., SPINK, A.: “Patterns of query reformulation during Web searching”. *J. Assoc. Inf. Sci. Technol.* 60(7), 1358–1371. 2009.

[26] KLEIN, G., MOON, B., AND HOFFMAN, R. F. “Making sense of sensemaking I: alternative perspectives”. *IEEE Intelligent Systems*, 21(4), pp. 70–73. 2006.

[27] KULES, BILL; SHNEIDERMAN, BEN. “Users can change their web search tactics: Design guidelines for categorized overviews”. *Information Processing & Management* 44(2):463-484. March 2008.

[28] LA ROSA M, DUMAS M, TER HOFSTEDE AHM, MENDLING J. “Configurable multi-perspective business process models”. *Inf Syst* 36(2):313–340. 2011.

[29] LEVIN, H., *Privatizing Education.* New York: Routledge. 2001.

[30] LI, G., HOU, Y. & WU, A. CHIN. “Fourth Industrial Revolution: technological drivers, impacts and coping methods”. *Chinese Geographical Science*. 27: 626. https://doi.org/10.1007/s11769-017-0890-x. 2017.

[31] LIAO, YONGXIN; LOURES, EDUARDO ROCHA; DESCHAMPS, FERNANDO;

et al. “The impact of the fourth industrial revolution: a cross-country/region comparison”. *Prod. vol.28*, e20180061. Epub January 15, 2018. https://dx.doi.org/10.1590/0103-6513.20180061. São Paulo, 2018.

[32] LIU, C., GWIZDKA, J., LIU, J., XU, T., & BELKIN, N. J. “Analysis and evaluation of query reformulations in different task types”. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1-9.ACM. 2010.

[33] LOEWENSTEIN, G. “The psychology of curiosity: a review and reinterpretation”. *Psychological Bulletin*, 116(1), pp. 75–98. 1994.

[34] LU, JIAHENG; et al. "String similarity measures and joins with synonyms". *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*: 373–384. 2013.

[35] M. TIBAU, S. W. M. SIQUEIRA, B. PEREIRA NUNES, M. BORTOLUZZI AND I. MARENZI, “Modeling Exploratory Search as a Knowledge-Intensive Process”. *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pp. 34-38, Mumbai, 2018.

[36] MARCHIONINI, G. “Foundations for personal information infrastructures: information-seeking knowledge, skills, and attitudes”. In: *Information Seeking in Electronic Environments*. (Ch. 4, pp. 61-75). New York NY: Cambridge University Press. 1995.

[37] MARJANOVIC O, FREEZE R. “Knowledge intensive business processes: theoretical foundations and research challenges”. *In: Proceedings of the 44th Hawaii international conference on system sciences.* HICSS ’11. 2011.

[38] MALHOTRA Y. “Integrating knowledge management technologies in organizational business processes: getting real time enterprises to deliver real business performance”. *JKnowl Manag* 9(1):7–28. 2005.

[39] MAYNARD, A. D. “Navigating the fourth industrial revolution”. *Nature Nanotechnology,* 10(12), 1005-1006. Nature Publishing Group. http://dx.doi.org/10.1038/nnano.2015.286. 2015.

[40] MOHAMED A. KHAMSI; WILLIAM A. KIRK. *The triangle inequality in ℝn. An introduction to metric spaces and fixed point theory.* Wiley-IEEE. ISBN 0-471-41825-0. 2001.

[41] NAVARRO, GONZALO. "A guided tour to approximate string matching". *ACM Computing Surveys*. 33 (1): 31–88. 2001.

[42] NETTO, J. M; FRANÇA, J.B.S.; BAIÃO, F.A.; SANTORO F.M. “A Notation for Knowledge-Intensive Processes”. *CSCWD 2013*: 190-195. 2013.

[43] NETTO, J. M; FRANÇA, J.B.S.; BAIÃO, F.A.; SANTORO F.M. “Evaluating KIPN for Modeling KIP”. N. Lohmann et al. (Eds.): *BPM 2013 Workshops*, LNBIP 171, pp. 549–561, 2014.

[44] NONAKA, IKUJIRO; TAKEUCHI, HIROTAKA. *The knowledge creating company: how Japanese companies create the dynamics of innovation*. New York: Oxford University Press. 1995.

[45] NUNES, B.P. et al. “Interlinking documents based on semantic graphs”. In: Watada, J et al. (Ed.). *17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2013*, Kitakyushu, Japan, 9-11 September 2013. (Procedia Computer Science, v.22), p. 231-240. Elsevier, 2013.

[46] NUNES, B.P. ET AL. “Can entities be friends?” In: G. Rizzo, P. Mendes, E. Charton, S. Hellmann, and A. Kalyanpur, editors, *Proceedings of the WoLE Workshop in conjuction with the 11th International Semantic Web Conference*, volume 906 of CEUR-WS.org, pages 45–57, Nov. 2012.

[47] OWEN, C. “Understanding Design Research. Toward an Achievement of Balance.” *Journal of the Japanese Society for the Science of Design* 5(2): 36-45. 1997.

[48] ÖZMUTLU, H.C., & ÇAVDUR, F. “Application of automatic topic identification on Excite Web search engine data logs”. *Information Processing & Management*, 41(5), 1243–1262. 2005.

[49] ÖZMUTLU, H.C., ÇAVDUR, F., SPINK, A.,& ÖZMUTLU, S. “Neural network applications for automatic new topic identification on Excite Web search engine data

logs". *Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology* (pp. 1–10). Medford, NJ: Information Today. 2004.

[50] ÖZMUTLU, S., ÖZMUTLU, H. C., & SPINK, A. "A day in the life of Web searching: an exploratory study". *Information Processing and Management*, 40(2), 319–345. 2004.

[51] PACE, S. "A grounded theory of the flow experiences of web users". *Int' Journal of Human–Computer Studies*, 60(3), pp. 327–363. 2004.

[52] PATRICIA YU. "The stratification of higher education in the USA and Taiwan: a comparative analysis of students' college-choice outcomes". *Compare: A Journal of Comparative and International Education* 0:0, pages 1-23. 2018.

[53] PENNIMAN, W.D. "A stochastic process analysis of online user behavior". In C.W. Husbands & R.L. Tighe (Eds.), *Proceedings of the 38th Annual Meeting of the American Society for Information Science (ASIS)* (pp. 147–148). Washington, DC: American Society for Information Science. October, 1975.

[54] PENNIMAN, W.D. "Modeling and analysis of online user behavior". In A.E. Petrarca, C.I. Taylor, & R.S. Kohns (Eds.), *Proceedings of the 45th Annual Meeting of the American Society for Information Science* (pp. 231–235). White Plains, NY: Knowledge Industry Publications. 1982.

[55] PIMENTEL, M. "A Computer Science Researcher Looking for a Way to ThinkDo the Research on Computers in Education (Um Pesquisador em Computação em Busca de um Modo de FazerPensar Pesquisas em Informática na Educação)". *Brazilian Journal of Computers in Education* (Revista Brasileira de Informática na Educação - RBIE). 2018.

[56] PIROLLI, P., AND CARD, S. K. "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis". *In Proceedings of the International Conference on Intelligence Analysis*. 2005.

[57] QIU, L. "Markov models of search state patterns in a hypertext information retrieval system". *Journal of the American Society for Information Science*, 44(7), 413–427. 1993.

[58] REICHERT M, WEBER B. *Enabling flexibility in processaware information systems—challenges, methods, technologies*. Springer, Berlin. 2012.

[59] RIEH, S. Y., COLLINS-THOMPSON, K., HANSEN, P., & LEE, H.-J. "Towards searching as a learning process: A review of current perspectives and future directions". *Journal of Information Science,* 42(1), 19–34. https://doi.org/10.1177/0165551515615841. 2016.

[60] RIEH, S.Y., & XIE, H. "Analysis of multiple query reformulations on the web: The interactive information retrieval context". *Information Processing & Management*, 42(3), 751–768. 2006.

[61] RUSSELL, D. M., STEFIK, M. J., PIROLLI, P., & CARD, S. K. "The cost structure of sensemaking". *Paper presented at the INTERCHI '93 Conference on Human Factors in Computing Systems*, Amsterdam. 1993.

[62] SCHWAB, K. *The Fourth Industrial Revolution*. In: World Economic Forum, 25–38. Switzerland, 2016.

[63] SIMON, H. *The Sciences of Artificial*. 3rd edn. MIT Press, Cambridge, MA. 1996.

[64] SPINK, A. "Toward a theoretical framework for information retrieval (IR) within an information seeking context". *In Proceedings of the 2nd international information seeking in context conference*, 12–15 August 1998, Sheffield. UK: University of Sheffield, Department of Information Studies. 1998.

[65] STONE, PETER; BROOKS, RODNEY; BRYNJOLFSSON, ERIK; et al. *Artificial Intelligence and Life in 2030.* In: One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report. Accessed: September 6, 2016.

[66] TAKEDA, H., VEERKAMP, P., TOMIYAMA, T., AND YOSHIKAWAM, H. "Modeling Design Processes." *AI Magazine Winter*: 37–48. 1990.

[67] TIBAU M., W. M. SIQUEIRA S., PEREIRA NUNES B., BORTOLUZZI M., MARENZI I., KEMKES P. "Investigating Users' Decision-Making Process While Searching Online and Their Shortcuts Towards Understanding". In: Hancke G.,

Spaniol M., Osathanunkul K., Unankard S., Klamma R. (eds) Advances in Web-Based Learning – ICWL 2018. ICWL 2018. *Lecture Notes in Computer Science*, vol 11007. Springer, Cham. 2018.

[68] TONON, ALBERTO; CATASTA, MICHELE; PROKOFYEV, ROMAN; el al. “Contextualized ranking of entity types based on knowledge graphs”. *Web Semantics: Science, Services and Agents on the World Wide Web* 37, 170-183. 2016.

[69] VACULIN R, HULL R, HEATH T, COCHRAN C, NIGAM A, SUKAVIRIYA P. “Declarative business artifact centric modeling of decision and knowledge intensive business processes”. *In: 15th IEEE international conference on enterprise distributed object computing* (EDOC 2011). 2011.

[70] VAISHNAVI, V., KUECHLER, W., AND PETTER, S. (Eds.). *Design Science Research in Information Systems*. January 20, 2004 (created in 2004 and updated until 2015 by Vaishnavi, V. and Kuechler, W.); last updated (by Vaishnavi, V. and Petter, S.), December 20, 2017. URL: http://www.desrist.org/design-research-in-information-systems/. 2004/17.

[71] VAKKARI, P. “Searching as learning: A systematization based on literature”. *Journal of Information Science* 42(1):7-18 · February 2016.

[72] VAKKARI P. “A theory of the task-based information retrieval process: A summary and generalization of a longitudinal study”. *Journal of Documentation*; 57(1): 44–60. 2001.

[73] VAKKARI, P. “Exploratory searching as conceptual exploration”. *Proceedings of the Fourth Workshop on Human-Computer Interaction and Information Retrieval*. 2010.

[74] WESKE M. *Business process management: concepts, languages, architectures*. Springer, Berlin. 2007.

[75] WHITE, RYEN W, AND ROTH, RESA A. “Exploratory search: beyond the query-response paradigm”. *Synthesis lectures on information concepts, retrieval, and services*, 1(1): 1-98. Print. 2009.

[76] WILDEMUTH, BARBARA M; FREUND, LUANNE. “Assigning search tasks designed to elicit exploratory search behaviors”. *Proc. of the Symposium on Human-Computer Interaction and Information Retrieval*. Cambridge, California, USA, October 04 - 04, 2012.