Otto-von-Guericke University Magdeburg

Department of Computer Science
Fakultät für Informatik
Data and Knowledge Engineering Group

Master Thesis

Adaptive and Explainable Search for Comics

Author:
Suraj Bangalore Shashidhar

June 29, 2023

Adviser

Supervisor
M.Sc. Sayantan Polley

Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Examiners

First Examiner
Prof. Dr.-Ing. Andreas Nürnberger

Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Second Examiner
Prof. Dr.-Ing. Sebastian Stober

Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Shashidhar, Suraj Bangalore:
Adaptive and Explainable Search for Comics
Master Thesis, Otto-von-Guericke University
Magdeburg, 2023.

Contents

Abstract

Comics are versatile art form that combines sequence of graphics and dialogue to spin a story. It is not only used for entertainment, but also for education manuals, operation manuals and others. They have held enormous influence over movie industry and offers diverse storytelling opportunities. However, current comic book search engines focus on metadata and popularity, neglecting content and search context, leading to dissatisfaction. To address this, we extract domain-specific facets like textual elements like genre, gender, topics, and visual facets like book cover, color, texture as well as lower level features to enhance the search experience and provide more tailored comic book retrieval based on content and user preferences.

Personalising search experience according to surrounding context information needs, requires understanding users intent. To be non-obtrusive, We develop a relevance feedback system that could adapt online, using user interactions. Adaptation is learnt through users mouse hover activity to discern between their interest and non-interests.

We address users concerns like “Why did we get these search results for the query?” , “How are two books from same search results similar or dissimilar?”, “What does search engine believe my likings are?”, “How did my previous activity impact current search results?” through providing global and local explanations, both visually and textually .

Due to non-availability of comic book retrieval benchmark datasets, We conduct an user study to evaluate performance of our algorithm and explanations against baseline search methods in the lab. Answering the above questions, we find that domain based facets retrieved more relevant books, local explanation is effective in comparing between books, user feedback improved the subsequent results significantly. Although explaining personalization was barely useful.

Acknowledgements

I am grateful to Fakultät für Informatik (FIN), Data and Knowledge Engineering group (DKE), Otto-von-Guericke Universität, Magdeburg, for allowing me to pursue the master thesis under them.

I would like to express my sincere gratitude to my supervisor, M.Sc. Sayantan Polley, for his helpful guidance, revision and encouragement. I am extremely grateful for Prof. Dr.-Ing. Andreas Nürnberger, Prof. Dr.-Ing. Sebastian Stober and M.Sc. Valerie Krug for reviewing and grading my thesis.

I would also like to express my special thanks to my friends and family for their support and companionship during this challenging but rewarding process.

1 Introduction and Motivation

1.1 Motivation

Comics are typically short books made up of sequential multimodal text and visual data. Industry dates back to the early 20th century and valued over $15 billion in 2022. United States is the largest market, followed by Europe and Asia [1]. The success of comic book movie adaptations, especially those from Marvel and DC, has greatly increased the appeal of comic book characters and stories. The market is undergoing a shift toward digital platforms, to reduce costs and improve access. Extensive comic book libraries are available through digital platforms like Comixology [2] and subscription programs like Marvel Unlimited [3]. Print sales are still more than 10 times the digital sales despite the rise of digital comics. Through social media, forums, and fan conventions, comic book fans participate in active online communities that promote conversation and creation. Hence it becomes important for users to have better ways of searching for comics in digital medium.

Users find books of their interest through book displays, manual search or word-of-mouth recommendation in book stores. Digital platforms use metadata based retriever (MBSE) or other users interests to recommend relevant book, without considering it's content. An attempt to retrieve individual similar comic book images [1] was made without considering complete book, leading to loss of information.

Apart from their read history, metadata or word-of-mouth, users may show interest in a comics due to its content. Given above drawbacks finding books with similar characters, style, and themes can be done with the use of book content [2]. Imagine looking for 'Plastic Man' and finding 'Batman' or 'DareDevil' which share similar genres [3], or books

[1] `https://shorturl.at/hkoMS`

[2] `https://shorturl.at/dfxUY`

[3] `https://www.marvel.com/unlimited`

with a similar visual style [4], which are not achievable with MBSE. When searching through MBSE, users already have specified outcomes in mind, this eliminates pleasant surprises like related works from unknown authors and prevents diversification [5]. Interpretable features(facets) could aid in diverse yet quality search results. This begs the question, "Are all users expecting the same results for their queries?" and "Can users make the search adapt by providing interacting with the system?".

Novices often go by simplistic facets such as book appearance, character name. Artists by layout structure, sketch technique, stylometery [6] [7]. Regular readers or fans by narrative and overall content. Against query book 'Wonder Woman', novice likely expects other 'Wonder Woman' or 'Justice League' books, fans could expect 'Miss America' or 'Sheena' that share an overall similar story line against query book. Original 'Wonder Woman' had some cheesy sketches, an artist searching using this could expect 'Famous Funnies' with zany sketches. Retriever should identify user's implicit thought processes and hone the search [8]. Furthermore, we attempt to make it interactive to help users take control of the system and consider their feedback. However there is a lack of transparency resulting in doubts in users mind, "Why did I get these search results?" and "Is it possible to understand reasons behind personalization?" [9]. By knowing such reasons, user can then control the system to fit their information needs.

Providing explanations that explains reason behind top-k results or options to compare different books from search results could reduce distrust in search results and increase satisfaction [10]. Hence, we make an attempt to create such a search engine that adapts to users needs and explains its results lucidly.

1.2 Aim of Thesis

Recent research in retrieving comic book images individually using its content have shown promising results [1]. Digitization of comics through extraction of its information has become easier due to powerful Machine Learning and Deep Learning approaches [11] [2] [7]. However, research on retrieval of comics based on domain based facets is relatively unexplored. This is worsened by the fact that, comics are usually copyrighted and access to large scale quality labelled datasets do not exist. To alleviate presence of scant datasets, we propose an online learning algorithm based on user interactions using triplet loss [12] [8] [13] called Triplet based Adaptation (TRB). We provide global explanations using facet's contribution towards top-k results in an attempt to answer relevance, local explanations to compare between books from search results. On top of that, we also provide textual explanations based on book cover, to explain relevance feedback adaptation and it's relationship with your screen activity.

Due to non-availability of comic book retrieval benchmarks, we conduct an user study in lab to evaluate our search engine on it's facets, personalization and explanations. We evaluate our system on explainable facet 'Usability' based on ISO/IEC 9126-4 standard [4] and also interactive precision measure. Usability is decided based on Effectiveness, Efficiency and Satisfaction. Tasks are tailored to make users use each of these elements, interact with the system and provide response that cater to the above metrics. In this thesis, we answer below research questions through evaluation from user study. We formulate our research questions as follows:

RQ 1 : What is the impact of domain based facets on search results quality?

RQ 2 : What is the impact of user-system interaction on satisfying user's search context?

RQ 3 : How does textual explanations generated from book cover help user understand personalization?

RQ 4 : How does comparison table explanation help user discern between books from search results?

1.3 Outline

The rest of this thesis is as follows: Chapter two 2 provides a background and related work on the key concepts relating to Cost Functions, Transformers, Relevance Feedback and Adaptability, Explainability, User Interface, User Study and Comics. Chapter three 3 outlines the specific material, methods, models, and metrics used in carrying out the experiments and analysis. Chapter four 4 provides the specific experimental setup, details of user study, and discussion of results. Finally, chapter five 5 summarizes the important results as well as provides ideas for future work.

[4] `https://www.usability.gov/what-and-why/usability-evaluation.html`

2
Background and Related Work

This section presents a high level overview of cost functions, transformers, information retrieval, relevance feedback and adaptability, explainability, user interfaces, and comics. A more detailed handling of the topics can be found in [14], [15], [16] and [17], [18] respectively. It concludes with user study metrics and comic book industry related concepts.

2.1 Cost Function

The performance of the model during training is assessed using a cost function, whereas evaluation metrics determines model performance on test dataset. Specific problem and ML algorithm employed will determine the cost function. Typically, cost functions are based on distance, impurity, and correlation. Regression algorithms usually employ correlation-based cost function. Decision tree techniques utilize impurity-based cost functions to determine appropriate split, clustering methods frequently use distance-based cost functions to assign cluster [19].

Triplet loss [20] is a distance based cost function used to learn representations. It shortens the gap between similar examples while increasing the distance between dissimilar examples, by a margin as shown in equation 2.1. Triplet loss based deep networks are also adopted to retrieve similar content [12].

$$\mathcal{L}_{\text{triplet}}(a,p,n) = \max\big(d(a,p) - d(a,n) + m, 0\big) \tag{2.1}$$

where a is the anchor, p is similar example, n is dissimilar example, m is the margin hyperparameter and d is the distance metric.

Discrete assessment measures like as precision, recall, accuracy, and F1 score are often used in classification problems [21]. F-Measure combines both precision and recall using harmonic mean as shown in equation 2.2.

$$\text{F-Measure (F1)} = \frac{2(\mathrm{P}\cdot\mathrm{R})}{\mathrm{P}+\mathrm{R}} \tag{2.2}$$

where *P* is precision, *R* is recall.

2.2 Transformers

Transformer [22] neural network architecture has become commonplace, especially for natural language processing (NLP) tasks and image computer vision (CV) tasks. Self-Attention is the bedrock of transformer architecture. Self-Attention of a particular position is assigned by relative importance given to that position by all the other positions. The output is then determined for each location in the sequence, using this weighted sum. Multiple heads of self-attention are used to capture different dependencies. Self-Attention mechanism can parallely process, and outperforms RNN encoder-decoder attention [23] networks. However, transformer can't handle arbitrary length sequence in one-shot and must be fed fixed length short sequences at a time.

2.2.1 Vision Transformers

Vision Transformer (ViT) [24] is a transformer based architecture that processes images. ViT turns an image into a sequence of tokens that can be acted on by transformer by chunking it into tiny areas and flattening them into a series of tokens. The transformer encoder layers then process these tokens to reveal the overall representations of the image. It has produced cutting edge results on a wide range of computer vision tasks, including object identification and image classification.

2.2.2 CLIP

Contrastive Language-Image Pre-Training (CLIP) [25] is a novel technique for learning combined representations of texts and images. CLIP is pre-trained on a huge dataset of pictures and their associated textual captions. Learning to match images with its captions, while contrasting them with other images and other captions creates a combined representation space for Image and Text data. Later, CLIP can be applied to a number of tasks such as image retrieval, captioning and classification that exploits combined space.

2.3 Information Retrieval

Users have differing needs from search. It can range from requiring exact facts about a person, specific answer to question to something more gray like doing a market analysis, getting to know more about the collection or inventing new scientific processes.

2.3.1 Lookup

A query is compared against collection of indexed documents using , which then returns the most relevant results based on an exact or approximate match of the query terms. Lookup-based search type aids user to seek straightforward information facts like location, name, year or specific answers to questions in a given document or collection of documents. Bing [5], Google [6] and other large web search engines use lookup based search for finding relevant web pages. [26]

2.3.2 Faceted

A facet can be described as an aspect or quality that can be used to categorize the document. Combining multiple facets (features), users may search through a collection of documents using the Faceted Information Retrieval method. This method gives people a simple way to browse huge document collection and find information that applies to their specific wants and needs. Faceted search interfaces are often used in digital libraries and e-commerce websites. For example: one can easily narrow down information related to specific author and their genre facets from collection of fiction books [27].

2.3.3 Interactive

Interactive Information Retrieval [IIR] allows users to interact with the search system and provide their feedback based on their perceived relevance of the retrieved documents, this effectively allows them to modify underlying search criteria. Users can gradually enhance the quality of search results by repeatedly changing their search

[5] https://www.bing.com/

[6] https://www.google.com/

phrase to represent their search context. IIR enables users to learn new information from retrieved results and supports limited exploration [28].

Related Work

Pausw et al. [29] introduces a playlist clustering system PATS for music. PATS dynamically clusters songs based on attribute-value similarity. The system's capacity to produce cohesive playlists with a wide range of songs was demonstrated in a user study that was used to assess its performance. The weights of the attributes are changed using an inductive learning technique, which also enhances subsequent playlists.

Nurnberger et al. [30] proposes SOM based document classification integrated with user feedback. Global weighing scheme is applied to move document close to target cluster. Inside large clusters, feature weights are adjusted locally to improve document class assignment. Information can be discovered outside of standard search, through interactive intent modeling, which improves user involvement with information retrieval systems.

To solve the vocabulary mismatch issue, Ruotsalo et al. [31] introduces the SciNet system represents probable intentions as directions in the information space. Users can offer input by adjusting the keywords on the Intent Radar, a radial display.

Ruotsalo et al. [32] looks at how faceted interactive query suggestions affect the efficiency and interest of exploratory information retrieval. Through user study, they demonstrate that dynamic faceted query suggestions increase recall without reducing precision, which enhances the overall effectiveness. It does not, however, greatly improve the performance of a single query. This suggests that in order to improve the efficiency of the entire exploratory search session, future research should concentrate on developing technologies that encourage guided situated navigation.

Medlar et al. [33] generates different queries that condense search results is used for exploratory search query suggestions. As users navigate down the search results page, they automatically create these alternative queries using a sequence-to-sequence autoencoder. Their interface's endless scroll feature enables constant updating of query suggestions and signal their relevance. User study conducted points out that query summaries are helpful for their exploratory literature searches in the scientific field.

2.3.4 Exploratory

Exploratory search is especially useful when users are not aware of their search needs or end goals and would require multiple ways to visualize, interact and retrieve information. Users would have a plethora of options to dig into document details, search, visualize clustering and tweak recommendations through interaction, thereby serendipitously assisting them in information discovery. In large and complex document collections, this exploration assists users in finding patterns, connections, and relationships. Exploratory search systems are often used in intelligence gathering, scientific process discovery and digital humanities [34].

Related Work

Exploratory search is made possible by the development of semantic technologies and the availability of Linked Open Data (LOD). Using music as an example domain, Dimitrova et al. [35] conducts a user research that investigates the difficulties and advantages of browsing linked semantic data sets. The results suggest areas for development to improve exploratory search with LOD and offer insights for assessing interactive exploration of linked semantic data.

Marie et al. [36] focuses on using semantic and linked data. In addition to offering a linked data-based exploratory search solution and presenting a cutting-edge software architecture for flexible querying, they also offer an overview of current methodologies and systems. For an exploration-optimized interface, the Discovery Hub web application was created and tested. In this study, a discussion about different approaches of evaluating exploratory search engines is presented. With the advancement of Linked Open Data (LOD), rich and interconnected datasets are now attainable.

Nguyen et al. [37] extensively uses LOD, for exploratory search using semantic technologies to provide more relevant outcomes in the movie domain. The main objective was to equip users to interact with the system.

Athukorala et al. [38] have researched, how information retrieval systems can differentiate between exploratory and lookup search behaviors by examining six different search tasks. They identified specific markers such as query length, scroll depth, and task completion time to identify exploratory search behavior, which has significant implications for the development of specialized and flexible IR systems.

Low et al.'s [39] proposal involves a map-based interface which enables users to browse semantically linked articles while exploring two decades worth of ISMIR publications [7]. Through dimensionality reduction, they align local maps to give users an experience of continuity during exploration. Evaluation showed that users can discover important connections between publications using this interface.

2.3.5 Comparison of Search

As we can see, each type of information retrieval has a high level of user engagement and real-time query refining. They vary, nonetheless, in terms of search broadness, query type, query complexity, and user interface UI. Faceted IR has structured well-defined questions and low to medium query complexity, whereas Interactive and Exploratory IR have broad ill-defined questions and high query complexity. While Faceted IR has a narrow to wide search scope, exploratory IR has a wide search scope, while Interactive IR has a narrow to wide search scope. Interactive and Faceted IR present results in a ranked list whereas Exploratory IR allows for multi-form visualization like ranked list or clustering that enables extensive investigation and knowledge enhancement on document collection as shown in table 2.1.

Table 2.1: Comparison of Interactive, Faceted, and Exploratory Information Retrieval

Feature	**Interactive IR**	**Faceted IR**	**Exploratory IR**
Query type	Open-ended	Structured	Open-ended
Query refinement	Real-time	Real-time	Real-time
Query complexity	High	Low to Medium	High
Search scope	Narrow	Narrow to Wide	Wide
User involvement	High	High	High
Result presentation	List	List	Visual
Result filtering	Limited	Comprehensive	Comprehensive
Result exploration	Limited	Limited	Comprehensive
Supports exploratory activities	No	No	Yes

2.3.6 Gaps

Search context of the user could acutely differ against models inner workings. User has varying expectations from the system depending on their tasks. This can change even during same search session. User who starts searching for fiction books in a digital

[7] `https://ismir.net/resources/related/`

library information retrieval system, typically looks into genre can narrow scope more by issuing queries related to author or year. System should be capable to understand users intent behind the search and satisfy their needs. Failing which leads to gaps such as Intent and Semantic Gap

Intent Gap

When a user's search intent does not sync with the search engine results, it leads to Intent Gap. For instance, a person looking for “bank" could be interested in learning more about the financial bank. The term “bank" could also appear in other semantically dissimilar terms such as power bank or river bank. The intent gap manifests itself when the search engine provides results pertaining to gadget or river. Techniques like query expansion, query classification, and intent modelling are used to alleviate the intent gap. Query Expansion uses additional information to enhance the user's query whilst Intent Modelling involves utilizing user's search history to predict their intent. Content based personalization is also utilized to find intent [40].

Semantic Gap

The semantic gap is the discrepancy between the representation of content by search engine and user's description of the content. For example: in an image retrieval search engine, user may try to find images with similar objects but search engine has only captured low level features such as color, texture and edges from images leading to mismatched expectations. User may get irrelevant images that share similar color profile but contains different objects. In other words, systems view of relevance and users view of relevance are out of sync leading to semantic gap. Relevance Feedback, domain relevant facets and re-ranking are utilized to lessen the semantic gap. Explainability can be used to explain reasoning behind search results for the query and help user to change their mental model to better suit the system's model [41].

2.4 Relevance Feedback and Adaptability

Relevance feedback is a technique that utilizes user feedback on the initially retrieved results for a specific query to hone the subsequent search process. Relevance feedback could make a system adaptable, if the user can directly change the parameters of the model to get desired output [13]. Furthermore, system can adapt to personalize user interests in a process called personalization. We interchangeably use personalization and adaptation throughout the thesis. Relevance feedback encompasses three distinct forms: explicit feedback, implicit feedback, and pseudo feedback. These variations capture different ways in which users express the relevance of the search results, enabling improvements in subsequent queries [42]. Rocchio relevance feedback is an important relevance feedback algorithm used in information retrieval to improve search results [43]. It calculates a new query vector based on relevant and non-relevant documents. The formula for Rocchio relevance feedback is 2.3:

$$\mathbf{Q}_{\text{new}} = \alpha \cdot \mathbf{Q}_{\text{old}} + \beta \cdot \frac{1}{|D_r|} \sum_{\mathbf{d} \in D_r} \mathbf{d} - \gamma \cdot \frac{1}{|D_{nr}|} \sum_{\mathbf{d} \in D_{nr}} \mathbf{d} \tag{2.3}$$

where:

- $\mathbf{Q}_{\text{new}}$ is the new query vector

- $\mathbf{Q}_{\text{old}}$ is the original query vector

- α, β, and γ are weighting factors

- D_r is the set of relevant documents

- D_{nr} is the set of non-relevant documents

- $\mathbf{d}$ represents a document vector

2.4.1 Explicit Relevance Feedback

Users may provide relevance feedback by explicitly marking results relevant or irrelevant. User's feedback is then further used to hone the ranking of the results in subsequent steps. Explicit feedback aims to fill semantic gap between user and system notion of relevance. Numerous information retrieval applications, including document retrieval and recommender systems, have extensively adopted explicit relevance feedback. [44]

2.4.2 Pseudo Relevance Feedback

Pseudo relevance feedback automatically identifies most relevant documents from the initial search results and extracts important terms or features from those documents. The query is then updated to include these extracted phrases for a more thorough search. Assuming the top k documents are already relevant, pseudo relevance feedback presupposes that their phrases are likewise relevant to the search context and overcomes the limitations of the initial query. This technique has been widely studied and applied in large scale information retrieval tasks, such as web search and document retrieval [44].

2.4.3 Implicit Relevance Feedback

There are other signs of user's notion of relevance. They may stay on a page for more time if they feel it suits their needs or only click on websites that have videos or continuously skip the search result page. Capturing such implicit signals and applying them as feedback to hone search is called implicit feedback. Implicit feedback is mostly non-invasive and can be derived from common user behaviour such as click-through rate, dwell time on a page, or query reformulations. Implicit feedback can directly impact search experience through re-ranking using captured user preferences, thereby improving perceived relevance. It has gained significant attention in the field of information retrieval and has been widely explored in various applications, including large scale web search and recommender systems [44].

Related Work

Reinforcement Learning based Feedback : Langford et al. [45] suggest the use of the epsilon-greedy algorithm to handle contextual multi-armed bandits. Interestingly, this method neither requires information about a time horizon nor does it sacrifice regret control for sample complexity bound; in fact, it achieves a regret scaling, where S represents the complexity term in supervised learning sample complexity bound.

In their study published as Chaudhuri et al. [46] , researchers focus on two forms of online learning situations with restricted feedback on top-k items and consequently feasible ranking strategies are offered to reduce regret while utilizing popular ranking measures and surrogates.

Li et al.'s [47] novel model presents an exploration function alongside its attractiveness function when studying online ranking algorithms; in other words, it utilizes an extended version of existing methods that takes into consideration a large number of items while producing better results compared to earlier approaches by relaxing the stringent probability considerations every time a user clicks on an item. user clicks on an item.

Click based Feedback : Joachims et al. [48] investigated reliability of implicit feedback on web search engines, using eye-tracking data and explicit relevance judgments to analyze user decision making patterns. According to their findings, user's clicking decisions are affected by relevance of search outcomes. However, adopting specific strategies could help align relative feedback signals with explicit judgments more accurately.

Another study by Joachims et al. [49] primarily focuses on analyzing the reliability of implicit feedback in web search using click-through data and query reformulations. By contrasting implicit feedback with manual relevance judgments through eye-tracking studies, they found that clicks were informative though biased. However, they also demonstrated that relative preferences derived from clicks were reasonably accurate within individual queries and across multiple sets of results in a query chain.

Chukiln et al. [50] proposed an inclusive method to transform any click model into an evaluation measure for integrating these two lines of research into offline metrics based on click models through transforming them into one shared evaluation framework. They compare derived model-based metrics with conventional measures highlighting stronger connections between model-based offline metrics and online outcomes particularly when assessing imprecise relevance assessments agreement between offline and online experimental results.

Distance based Feedback : Buckley et al. [51] investigated approaches for learning weights using relevance data from a learning set of texts. Improving upon Rocchio algorithm [43], the Dynamic Feedback Optimisation technique dynamically improves query weights. When tested against the test set, the optimised query performs 10-15% better than the original. Discrepancy between features and concepts is overlooked along with perception subjectivity.

Rui et al. [52] presents a relevance feedback method that takes these qualities into account and tries to fill semantic gaps. Weights are dynamically updated based on user feedback to account for high-level questions and subjective perception. Two weights are maintained, one for Intra Feature and another for Inter Feature. The outcomes of the experiments show increased retrieval accuracy and decreased user effort.

Stober et al. [13] creates a music explorer system 'BeatlesExplorer' that enables user to adaptively explore and search the collection. It is based on growing self organizing maps (GSOM) . They evaluated various optimization algorithms such as gradient descent, max margin classifier and quadratic programming approaches for constraint violations. Gradient descent (GD) based approach gets stuck in local minima and doesn't satisfy few constraints, but can still work if constraints are violated. Adaptation based on triplet loss can be defined as triplet based adaptivity (TBA).

The unique method of Deep Triplet Quantization (DTQ), which combines deep learning and quantization Liu et al. [53] compacts binary codes through triplet selection approach and a quantization loss. In image retrieval benchmarks MS-COCO [54] and CIFAR-10 [55], experimental data show that DTQ performs better than the state-of-the-art hashing techniques.

Chittajallu et al. [56] proposes a video frame retrieval system with XAI techniques on minimally invasive surgery (MIS) videos. They enhance the retrieval process by incorporating relevance feedback and visual explanations, ultimately achieving positive outcomes on the Cholec-80 [57] dataset.

Yang et al. [58] proposes an end-to-end framework, that facilitates facial image retrieval in forensic investigations. Their model is unique in its iterative and interactive approach involving relevance feedback, and achieves high performance without additional annotations through tests on the CelebA dataset [59].

Wang et al. [60] creates Convolution Neural Network CNN based semantic re-ranking system for improving sketch-based image retrieval (SBIR). The system utilizes dual CNN models trained for sketch and image classification to capture semantic information. They re-rank the initial result set using a category similarity measurement. It also attains better precision performance across various SBIR datasets compared to others.

Jia et al. [61] proposes an online learning to rank (OL2R) approach that estimates a pairwise learning to rank model. The approach partitions and ranks candidate documents based on the model's confidence, focusing exploration on uncertain pairs of documents. In this way they plan to bridge offline and online learning. Experimental comparisons with OL2R baselines on benchmark datasets validate the effectiveness of the proposed approach.

2.5 Explainable Artificial Intelligence

For over several decades now researchers have been exploring Explainable Artificial Intelligence solutions (XAI), few beginning studies offering insights into these concepts [62]. Yet recent trends within Machine Learning (ML) are driven by the availability of larger datasets and increased computing power resulting in ML systems that now supersede human performance capacities across varying interfaces - an instance being AlphaGo's victory over Go champion Lee Sedol [63]. While earlier versions of the AI systems only required simple algorithm executions explaining their every move in detail wasn't sought after however complex algorithms like deep neural networks emerged posing challenges requiring interpretability by users. XAI consequently proved vital in engineering solutions necessary within different domains ranging from factory automation through banking infrastructure supporting critical lifesaving applications. Ultimately, XAI algorithms are geared towards providing explanations to AI predictions. It's worth noting that the discipline of XAI is not entirely comparative or reliant upon Artificial Intelligence (AI) integrations rather it serves as part of a broad interface responsible for improving effective human-Agent interrelationships compared to most traditional algorithms. As AI systems became increasingly prevalent in industries such as factory automation, banking, healthcare, defense, law, and critical infrastructure, the necessity for XAI became more evident [10]. However, with the rise of complex algorithms, including deep neural networks, interpretability became a challenge. Therefore, XAI aims to deliver efficient explanations of its decisions, taking into consideration the intended recipient of the explanation.

There is also small difference between explainability and interpretability. However we use them interchangeably for explainable search in current context. Explainability attempts to explain models decision making process to humans, whereas Interpretability does not deep dive into models working but aims to help humans understand cause and effect of model against input. We can categorize explainability into below categories.

2.5.1 Categorization of XAI

Ante-Hoc versus Post-Hoc

Ante-Hoc explanations refers to simplistic class of models such as Decision Trees, Linear Regression whose output are self explanatory and can be traced back against input. For example: Linear Regression models feature weights can tell user about contribution of feature into making the decision. Decision tree model can provide specific rules applied to reach the output. Post-Hoc explanations are used to explain black box models such as Neural Networks through other special AI techniques and supplementary models [64].

Model Specific versus Model Agnostic

Model Specific explanation techniques exploit algorithm's inner workings and provide explanation that is valid only for that particular model. For example: Tree Shap [65] and Integrated Gradients [66] are specifically used to debug Decision Trees and Neural Networks. Model Agnostic techniques treat model as black box and does not use any algorithm of model to generate explanations. For example: LIME [67], SHAP [65], Feature Importance [68] are frequently used techniques to generate explanations across multiple models and compare them.

Static versus Dynamic

Static Explanation provides terse reasons which cannot be explored more by the user. Dynamic explanation allows user to ask for more information about explanation. They can also interact and change the explanation and evaluate cause-effect relationship between explanation and system [69].

Local versus Global

Local Explanation deals with explaining individual output. For example: in a book genre classifier, a book can be classified as Literary, Detective, Romance. Local explanation could tell user the important words that lead to Sherlock Holmes being classified as Detective book [9]. Global Explanation gives a summary about the model's decision making process to the user. For example: in a Cardiac Arrest Prediction system that uses age, gender, cholesterol level, bmi as features, Global explanation can tell user that cholesterol level and age were important to predict cardiac arrest for the dataset and rest were not so important [9].

2.5.2 Explainable Search

Explainable AI (XAI) approaches differ depending on the scenario, with classification settings emphasizing add-on methods like SHAP, LIME and Feature Importance for explaining classification decisions. In the domain of information retrieval (IR) systems, the challenge lies in explaining the process of generating rankings or relevance of search results. Rankings are typically derived using similarity measures to estimate relevance, particularly in ad-hoc text queries. Relevance is also influenced by factors like context, application scenario, and user search context along with similarity. Relevance can be elucidated globally through top-k search results or locally individual comparison between query and individual result book. For example: in comic book searching, the goal is to elucidate facets responsible for top-k results, considering domain specific aspects like genre, gender and book cover. User can have some search context such as preference to genre over other facets which needs to be accommodated into explanation. However, obtaining ground truth relevance and explanations poses challenges, especially in the XAI setting where ground truth is limited [9].

Related Work

Post-Hoc : Qiao et al. [70] interprets a BERT-based ranking model using feature ablation. They compute token importance scores by comparing the ranking scores of a document with and without the removal of randomly chosen input tokens. They discover that the ranking score heavily relies on a small number of tokens, which are often exact match terms from the query and terms in close semantic context.

Singh et al. [71] creates explainable search engine EXS, that leverages feature attribution to give users explanations. It extends LIME, originally created for classifiers, to rankers by converting query document pair scores into class probabilities, treating document ranking as a classification issue. The three main issues of document relevancy, ranking, and query intent are addressed by EXS.

Polley et al. [72] proposes evidence based explainable search against EXS. In a user study, ExDocS was assessed as having higher interpretability whereas EXS had comparable completeness and transparency scores with a drop in performance against EXS.

Polley et al. [73] proposes an explainable AI-driven image retrieval system X-Vision, that produces textual and visual explanations. It bridges semantic gap in content-based image retrieval and offers re-rankings based on rules to boost retrieval efficiency. A re-ranking system that produces explanations, improving user confidence and comprehension of similarity in image search, is evaluated on PASCAL VOC data reveals enhanced retrieval performance above baselines.

Polley et al. [74] presents SIMFIC 2.0, an Explainable Search engine to find similar fiction books. It focuses on assessing the trustworthiness of the offered explanations as well as retrieval performance. The system makes use of domain specific interpretable features and offers both global and local explanations. To assess trustworthiness and ranking performance, user study and click analysis are carried out. To assess user attention to explanation elements, eye tracking is used. The findings of the system's trust are statistically significant, according to early experiments.

Singh et al. [75] suggest an approach for producing explanations that involves transforming a black-box ranker into an interpretable global surrogate model. The surrogate model imitates the predictions of the black-box model and acts on interpretable features. They carry out experiments to show that, despite the fact that an interpretable ranker can be trained for some query locales, there are restrictions on explaining all localities of a complicated model. In such circumstances, local surrogate models may be beneficial.

Singh et al. [76] introduces a greedy search-based method for locating explanations in Learning to Rank (LTR) models. In order to enhance both validity and completeness, their methodology focuses on selecting a subset of explanatory features. Validity is the assurance that the chosen features have the necessary predictive power to accurately duplicate the original result ranking.

Contrary to methods that are model agnostic, model introspective feature attribution techniques demand access to the model's internal workings. These techniques compute feature importance scores using gradients or other model features. A model-introspective technique for feature attribution in LTR models is suggested by Purpura et al [77]. To determine crucial features and produce saliency maps, they employ gradient-based feature attribution. By calculating the frequency of importance throughout the saliency maps, feature groups are further selected based on thresholded importance values.

Fernando et al. [78] uses DeepSHAP, a synthesis of SHAP and DeepLIFT from Lundberg et al. [79] in neural retrieval models. They look at how sensitive explanations are to various baseline input document selections. Based on various baseline inputs, comparisons between DeepSHAP and LIME show differing degrees of overlap in crucial aspects.

Explainable by Design : Khattab et al. [80] measures document relevance by computing the sum of maximum cosine similarity score between each query token and document token. They focus on capturing semantic similarity between the query and document, considering documents more relevant if they contain terms that are semantically closer to the query.

Lucchese et al. [81] proposes ILMART, a LambdaMART-based [82] interpretable Learning to Rank model. ILMART achieves a compromise between ranking quality and model complexity by efficiently training ranking models using a small and controlled number of pairwise feature interactions. ILMART beats current state-of-the-art interpretable ranking algorithms, according to experimental data, which show a considerable gain in normalized discounted cumulative gain (nDCG) of up to 8%.

2.6 User Interface

User Interface refers to visual presentation of the system to the end user through which user could interact with the system to fulfill their needs. For IR systems, user Interface comprise of query form display, search results presentation and supplementary data presentation. Information Retrieval systems tend to mix and match above interfaces according to their needs.

2.6.1 Query Form Display

On the classic search interface, users place their search words into a text entry form. Because form width and query duration are linked, longer queries may be discouraged in smaller forms while longer searches may be promoted in larger forms. Because they incorporate numerous components, some forms allow users to enter a wide query and then narrow it down using filters. These forms can store previously input data, allowing users to establish criteria for subsequent search searches. Furthermore, search forms usually give recommendations through highlighted language to assist visitors determine what type of information to enter. As users interact with the form, the highlighted text disappears, allowing them to input their search phrases [83].

2.6.2 Search Results Presentation

List Based

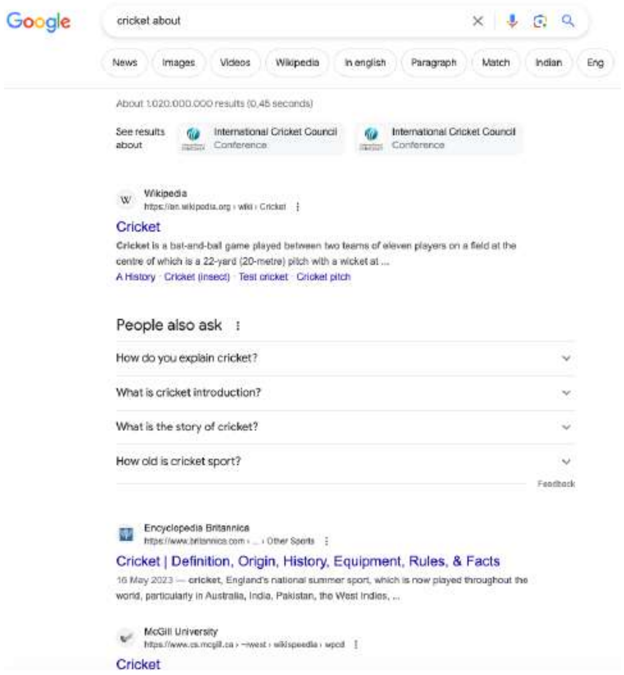
Search results are ordered by its relevance with query in the form of list. This layout is easy for the user to pick most relevant documents among top-k results as shown in fig 2.1. List based UI creates bias in user about the ranking of document and can influence them to view only few documents among top-k. User can miss out any relevant document at end of the list. Furthermore, user search context may vary and they may not be looking for only relevance scores but also how results are related to one another or query to know more about the corpus. [84]

Graph Based

Graph based user interface allows users to drill up or down and find relevant documents linked to their interest. Users can find relationship of the interested document to its neighbours and also with entire corpus as shown in fig 2.2. Graph based UI also attempts to remove ranking bias which increases diverse results. The links between documents can be formed through hyperlinks, relevance score, metadata and other domain specific data, which allows users to truly explore the corpus. However displaying a large corpus and relationships can overwhelm the user for which they need to

[8] Source: `https://www.google.com/`

Figure 2.1: List based UI from google search[8]



cricket about

About 1.020.000.000 results (0,45 seconds)

See results about International Cricket Council Conference International Cricket Council Conference

Wikipedia
https://en.wikipedia.org › wiki › Cricket

Cricket

Cricket is a bat-and-ball game played between two teams of eleven players on a field at the centre of which is a 22-yard (20-metre) pitch with a wicket at ...

A History · Cricket (insect) · Test cricket · Cricket pitch

People also ask

How do you explain cricket?

What is cricket introduction?

What is the story of cricket?

How old is cricket sport?

Encyclopedia Britannica
https://www.britannica.com › ... › Other Sports

Cricket | Definition, Origin, History, Equipment, Rules, & Facts

16 May 2023 — cricket, England's national summer sport, which is now played throughout the world, particularly in Australia, India, Pakistan, the West Indies, ...

McGill University
https://www.cs.mcgill.ca › ~west › wikispeedia › wpcd

Cricket

have dynamic zooming over the UI. Graph based UI may not be suited for data which is in the form of hub and spoke and few documents end up with lot of connections and others with none. [84]

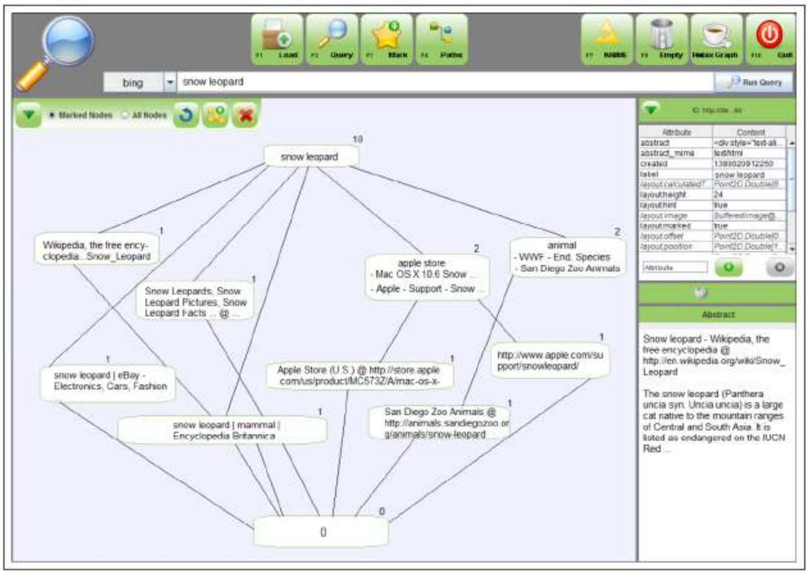
Map Based

Map based UI presents results preferably on a 2-d map though which users can truly start exploring the corpus. Map based UI enables user to perform both search and navigation tasks seamlessly where users can dynamically zoom in or out of local neighbourhoods to global as shown in fig 2.3. They also helps to alleviate bias with rankings. Presence of large hubs can still impact Map based UI though filter bubbles. In such bubbles, user will be bombarded with same search results again and again and would be unable to get out. User would need to zoom out and get out of the bubble manually [84].

2.6.3 Supplementary Data

Supplementary data refers to any additional information such as metadata, summaries or explanations about the search result documents that could help users to know more

Figure 2.2: Graph based UI from a prototype taken from Low et al. [84]


F1 Load
F2 Query
F3 Mark
F4 Paths
F7 NABS
F9 Empty
Relax Graph
F10 Quit
bing
Run Query
Marked Nodes
All Nodes
10
snow leopard
1
Wikipedia, the free ency-clopedia..Snow_Leopard
1
Snow Leopards, Snow Leopard Pictures, Snow Leopard Facts ... @ ...
2
apple store
- Mac OS X 10.6 Snow ...
- Apple - Support - Snow ...
2
animal
- WWF - End. Species
- San Diego Zoo Animals
1
http://www.apple.com/support/snowleopard/
1
snow leopard | eBay - Electronics, Cars, Fashion
1
Apple Store (U.S.) @ http://store.apple.com/us/product/MC573Z/A/mac-os-x-
1
snow leopard | mammal | Encyclopedia Britannica
1
San Diego Zoo Animals @ http://animals.sandiegozoo.org/animals/snow-leopard
0
0
ID: http://de...le/
Attribute
Abstract

snow leopard

Attribute	Content
abstract	<div style="text-ali...
abstract_mime	text/html
created	1380020912250
label	snow leopard
layout.calculated?	Point2D.Double[8...
layout.height	24
layout.hint	true
layout.image	BufferedImage@...
layout.marked	true
layout.offset	Point2D.Double[0...
layout.position	Point2D.Double[1...

Snow leopard - Wikipedia, the free encyclopedia @ http://en.wikipedia.org/wiki/Snow_Leopard

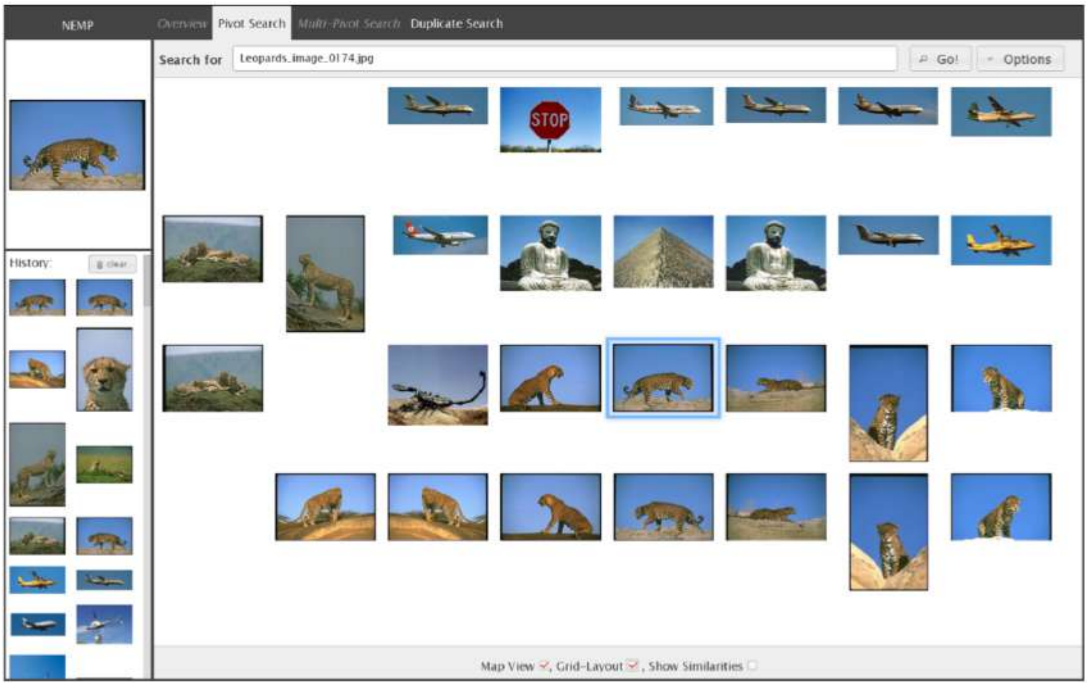
The snow leopard (Panthera uncia syn. Uncia uncia) is a large cat native to the mountain ranges of Central and South Asia. It is listed as endangered on the IUCN Red ...

about their documents. Typically important topics related to document are usually provided right below its title that could help users to choose their next document. If there are more information then it is typically presented in a side panel. Supplementary information could be presented to user on demand by hover or double click to stop overwhelming the user.

2.7 User Study

User Study in the context of information retrieval refers to, investigations into human - search engine interaction around many measures. These studies consider several measures such as accuracy, recall, precision, usability, ease of use, satisfaction and others. With the advancement of information usage leading to varied challenges and approaches in user-oriented research, incorporating humans themselves. Modern research now includes a broader spectrum of studies concerning them. One can classify the different types along two lines that stress either human-related traits or system characteristics. This vast range lets researchers look into numerous subjects pertaining to designing search systems suited for user's preferences or examining their performance metrics. Ultimately such studies benefit by addressing problems directly or indirectly tied with developing or evaluating these types of systems [85].

Figure 2.3: Map based UI from a prototype taken from Low et al. [84]



Search for Leopards_image_0174.jpg

Map View ✓, Grid-Layout ✓, Show Similarities □

2.7.1 Composition of User Study

Research Question

Research question is the initial starting point of any user study. It broadly defines the gaps being addressed in the research area. Research question do not directly make comparative claims about the two systems like Is system X is more user friendly than system B, but What is the impact of new feature on user friendliness on systems [85]

Hypothesis

Hypotheses are derived from research questions and clarify predicted relationships among those concepts. The ideas indicated in those investigations are finally represented by diverse variables. Alternative and null hypotheses are two different categories of hypotheses. The researcher's affirmation of the expected relationship between the concepts under investigation often referred to as the alternative hypothesis. As they offer alternatives to the null hypothesis, which asserts that there is no relationship or difference. By default, the null hypothesis is accepted, hence it is up to the researcher prove that the relationship exist. For example: if a new feature allows user satisfaction to improve the system, hypothesis can be formulated as, is system X provides more satisfaction to users than system Y. Hypothesis are usually validated by statistical testing between groups[85].

Variables

Concepts are represented by variables, which also include the methods used by researchers to define, observe, and quantify these ideas. Research related notions like relevancy, effectiveness, and satisfaction are expressed using variables. Mostly, variables needs to be redefined or conceptualized according to research question. For example: notion of relevance is different between two systems. In research studies, variables can be classified into different categories based on their roles. Independent variables are considered as the causes, while dependent variables are the effects. During experiments, researchers have control over manipulating the independent variable, such as instructing participants to use specific systems or assigning them to particular conditions. On the other hand, quasi-independent variables are factors that can influence the outcome measure but are not directly manipulated by the researcher. An example of a quasi-independent variable is gender, where researchers may want to examine differences between males and females in their usage of experimental and baseline information retrieval systems. However, the researcher cannot manipulate an individual's gender. Confounding variables, also known as confounds, are variables that impact the independent or dependent variable but are not controlled by the researcher. These variables may be discovered during or after the study, and if identified beforehand, researchers can account for their effects to ensure accurate results. For instance, in a study investigating the impact of search experience on information retrieval system success, researchers would need to ensure equal representation of participants with both high and low search experience across the tested systems [85].

Design

Evaluating hypothesis could have few bias that could poison user study. Evaluating task 1 first could help users performance in task 2. However vice-versa may not be true. Similarly evaluating System X before than System Y could make user rate System Y better than X. Factorial designs are used to evaluate hypothesis that could contain multiple variables. Subjects could evaluate multiple systems or can evaluate only one system. The tasks can be rotated and order of systems can be switched from user to user to reduce bias. This is called Greco-Latin square design [85].

Measures

Measures are used to operationalize the variable. Variable must be exhaustive in range of options that users should choose. Options must be unique and must not be overlapping with one another. If possible, make user understand about various options through examples before user study [85].

2.7.2 Categorization of Measures

Contextual

Contextual measures describe information about the subject that is user. For a few user studies location, age, gender, familiarity with the topic, incentives matters a lot. For example: a children search engine should be evaluated by filtering on age. Incentives provided to user also determine their interest in user study [85].

Interaction

Interactive measurements involve quantum of interaction between user-system. Collecting data on documents seen, time stayed on search results page, click rate and other user habitual data. Typically user interaction is logged and later it is evaluated in conjunction with others. Eye-tracker is also employed to measure attractiveness or distress to complement quantitative data [85].

Performance

Performance based metrics gauge effectiveness of the interaction include the number of relevant documents mentioned. Precision, Recall, Discounted Cumulative Gain and other metrics are used to measure performance of the system. Interactive Precision and Recall are especially used when user interacts with a system and saves certain document as normal Precision and Recall considers all documents for calculation. [85].

Usability

Usability of system is usually subjective. Users are elicited to self-report measures under the umbrella of usability, elicits individuals views and sentiments about the system [85]. It is further sub-divided into Effectiveness, Efficiency and Satisfaction. Effectiveness describes task performance rate in the system, whereas efficiency represents time required to perform the task. Satisfaction is subjective to the user. Effectiveness and efficiency are also considered as part of performance measures in some cases.

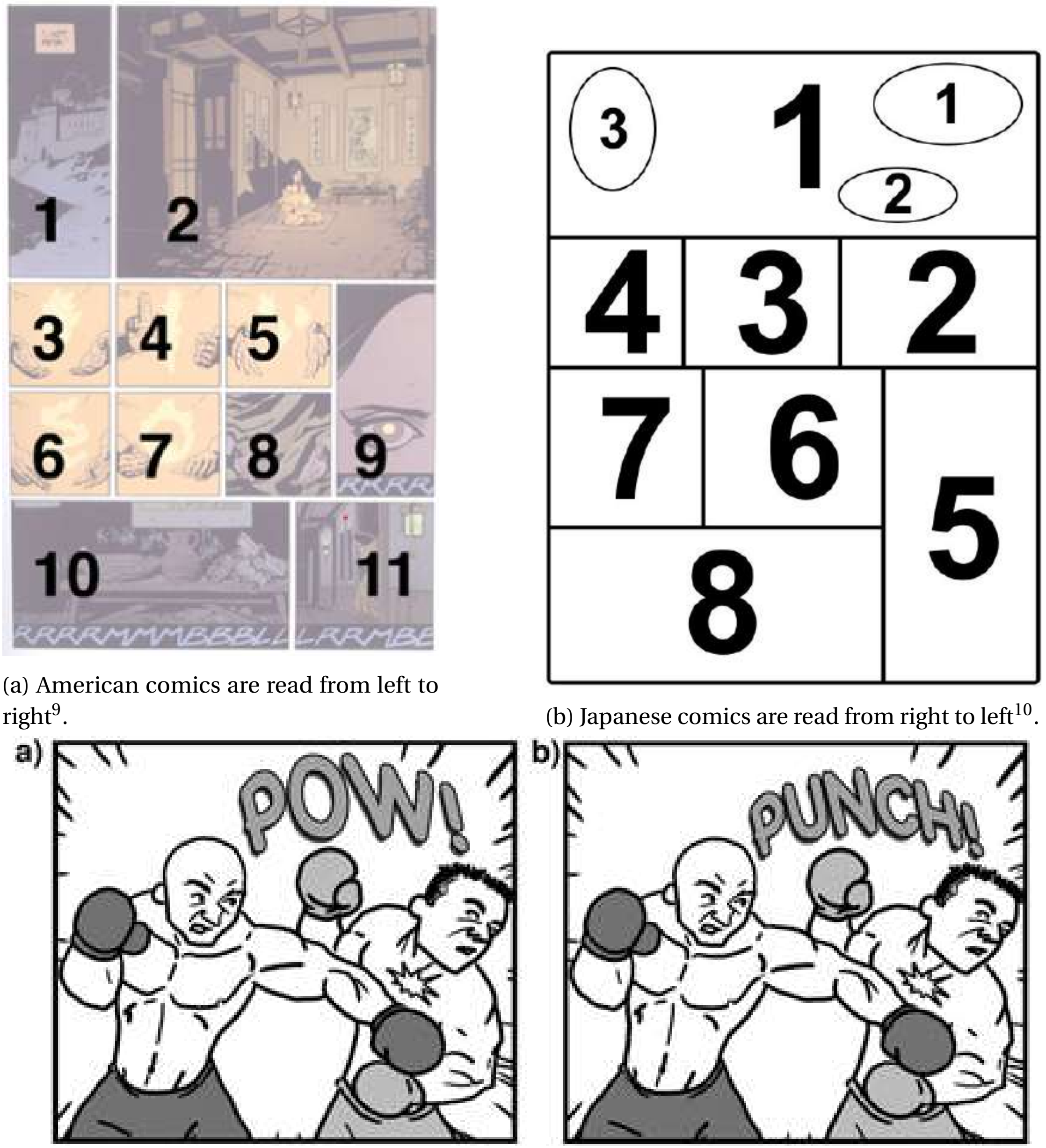
2.8 Comics

Comic books are a type of sequential art that uses words and images to convey stories. Speckled with vivid illustrations alongside concise dialogues in a serialized format. Individual issues are bundled into larger volumes or graphic novels. Visual storytelling is fundamental to comic books. Every image is an integral part of the narrative paired with text. This powerful combination generates an immersive experience unique to comic books. Fantasy and science fiction themes primarily inspire comic story lines leading to characters with exceptional abilities living in extraordinary worlds- this explains fairly why certain stories appeal to fans of all age groups [86]. Unlike other literature forms such as fiction or graphic novels, comic writings rely more on imagery than written prose. Story arcs that unfold across multiple volumes, readers intensely wait out them to unravel.

Notably, popular culture has embraced comic's dominant figures like Justice League's Green Lantern, Iron Man and Captain America that transcend print boundaries fitting into contemporary media such as television shows or games all underscoring their significance as valued pieces of literature [86].

2.8.1 Comics Culture

Cohn et al. [87] describes cultural variation in comics, especially due to geographies. For instance, American, Japanese, and Korean comics all display different comic book cultures that each have their own unique traits. The reading order is one notable distinction. While Japanese comics, known as manga, are typically read from right to left, American comics are read from left to right, top to bottom. Manhwa, or Korean comics, often follow the same left-to-right reading convention as American comics. The usage of onomatopoeia is another noteworthy feature. Onomatopoeic emotions and the incorporation of dynamic, visually striking sound effects into the artwork play a big part in Japanese comics. Onomatopoeia is also used in Korean manhwa, however it is used less frequently than it is in Japanese comics. On the other hand, American comics typically use more traditional sound effects [88]. This is illustrated in fig 2.4.



(a) American comics are read from left to right[9].

(b) Japanese comics are read from right to left[10].

(c) Changes in onomatopoeia between American and Japanese comics[11].

Figure 2.4: Illustration of diversity in comic culture across geographies.

[9] `https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00186/full`

[10] `https://commons.wikimedia.org/wiki/File:Manga_reading_direction.svg`

[11] `https://www.degruyter.com/document/doi/10.1515/mc-2016-0017/html`

2.8.2 Panel

Panels are the most basic semantic unit of the comic book. Each pane consists of a visual scene, dialogue text and expressions. In essence, panel captures an important snapshot of the scene. Panels are usually of rectangular shaped and are separated from each other by a white border called gutter as shown in fig 2.5. Panels could be arbitrary polygon shaped or have curved borders on which bounding boxes could not be drawn[89].

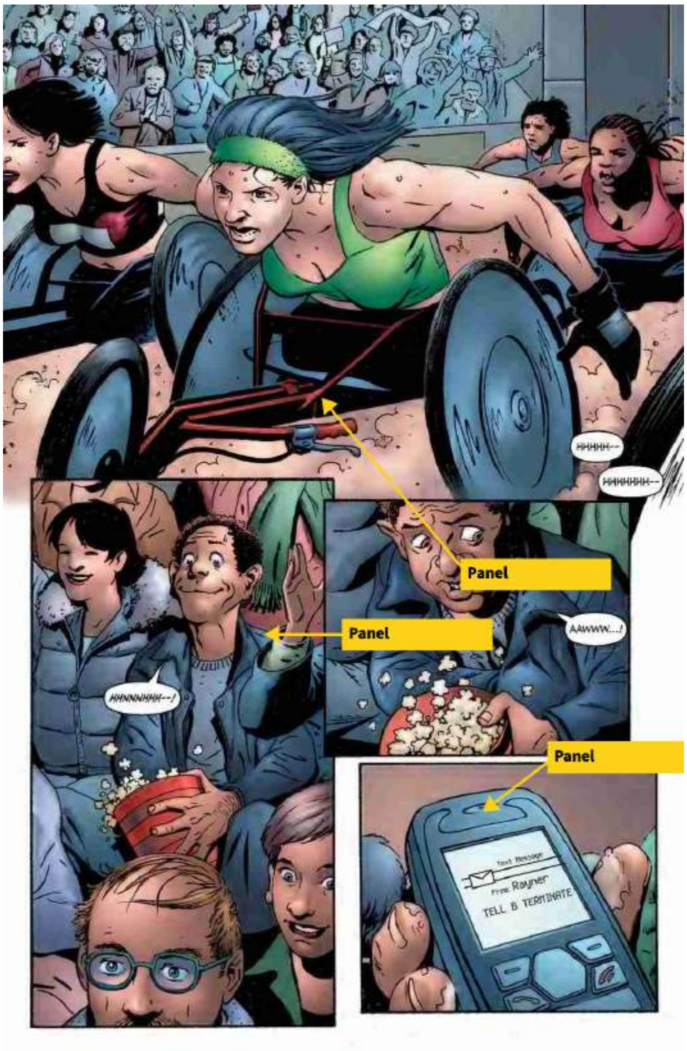


Figure 2.5: Panels with no box and overlaps.[13]

[13] https://www.dynamite.com/htmlfiles/viewProduct.html?PRO=C725130098443

Gutter

The basic function of gutter is to separate panels. Artists cleverly use gutter space and shape to change tempo of the scene. Gutter shape and size also used to create curiosity. Last panel from page could be removed purposefully to allow user to fill in their thoughts and perspective as shown in fig 2.6. Gutters can also be removed or faded to provide continuity or fast paced action between panels [90].

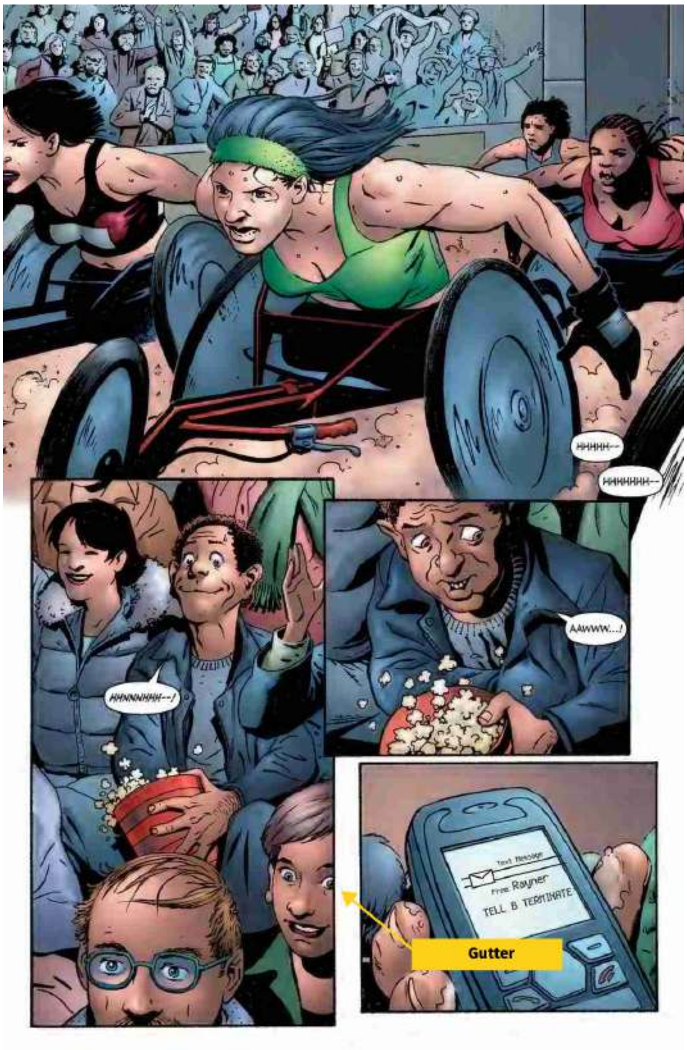


Figure 2.6: Uneven Gutter Spacing prompts user to fill in their thoughts and alter time transition [14].

[14] https://www.dynamite.com/htmlfiles/viewProduct.html?PRO=C725130098443

Related Work

Extracting panels are not only useful for digitization but also for displaying them on mobile screens In et al. [91]. Simplistic panels with 4 edge polygons and clear gutter can be extracted with the help of line-cutting algorithm Li et al. [92]. However, they do not consider free space between panels and connected component approaches have been used used to extract panels Ho et al. [93]. Connected component approaches could sometimes combine multiple panels into one if they are close enough. Iyyer et al. [11] used CNN based Fast-R CNN [94] to extract panels with better performance. They created a labelled dataset of 500 pages with bounding box information and were able to extract rectangular panels. However, this approach fails for circular and non-polygon panels. Laubrock et al. [95] used UNet segmentation [96] based CNN to successfully extract panels from GNC corpus [97] with state-of-the-art F1 Score.

2.8.3 Page Layout

Artist could choose to lay panels by varying its shape, size and gutter to convey their story. Panel sequencing is akin to composing section in literary text. Artists arrange panels on a page in a way, that is aesthetically pleasing and harmonious. Symmetrical composition is often used to create a sense of order, stability whilst non-symmetrical arrangement symbolize dynamism. Typical panel types are illustrated in fig 2.7 are as follows.

1. **Regular :** Symmetrical composition is often used to create a sense of order, scene continuity and stability. Grid layout allows user to smoothly glide or skim through the scene [87].
2. **Blockage :** Blockage panel is a panel that covers a large area of the page and dominates the composition. It is often used to emphasize a significant moment or action in the comic [87].
3. **Separation :** Panels are separated when they belong to two different events or there is a time gap between the two panels.
4. **Overlap :** Overlap panels are used when author wants to establish continuity and linkage between two panels [87].
5. **Staggering :** Staggering is typically used to disturb the flow of the user, usually to pick up or slow down pace of the story [87].

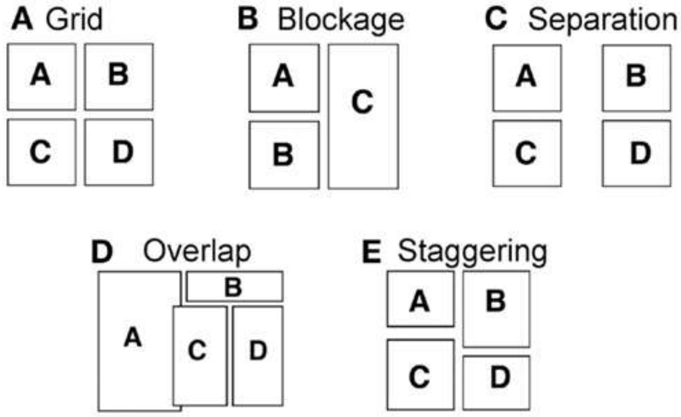


Figure 2.7: Different types of page layouts [87]

Related Work

Yusuf et al. [98] conducts a user study with children to understand their method of reading graphic novels. They use an eye-tracker to measure fixation of children on chosen comics. Children found difficult to navigate page layouts that have blockage or staggering. Wildfeuer et al. [99] examines the page layout of first-aid instruction comics demonstrating the Heimlich manoeuvre. Through a multi-level annotation scheme, they analyze similarities between the instruction graphics and typical comic book layouts. They find that although many elements between comic page layouts and instruction manuals are similar, instruction manuals tend to follow simplistic grid or list based presentation as they do not represent art. Bares et al. [100] creates an automated system to generate comic book page layouts that follow specific temporal order. The system takes into account the reading time, reading path, and style guides. The goal is to ensure that the size of each panel corresponds to the amount of narrative time or reader engagement. Pederson et al. [101] analyzed American superhero comics over an 80 year period to examine changes in page layouts. By examining a corpus of 40 comics from 1940 to 2014, the researchers found a decrease in the use of regular layouts and an increase in irregular layouts such as bleeds. Artists began treating entire page as a semantic unit rather than individual panel through splash panels or full page bleeds.

2.8.4 Characters

Comic books are heavily based on its central characters, detecting such important characters and tracking them through the story can tell us their overall appearance ratio in each book and also help us determine antagonist and protagonist in some scenarios. Detecting and tracking them is more challenging task than panel extraction. To begin with comic characters are more often non-human For example: Martian Manhunter and also non-humaniod For example: Parallax. Characters are in a constant state of action throughout the book hence their poses and cuts change from pane to pane. Super hero characters often leave out their masks and resume their normal life identity, there are no mechanisms to detect a character when they are not in their super hero identity. More often at each era characters are recast with different ethnicities and genders. A human would instantly know about the superhero regardless of gender and ethnicity but for a machine it would be difficult as illustrated by below picture of green lantern corps 2.8.

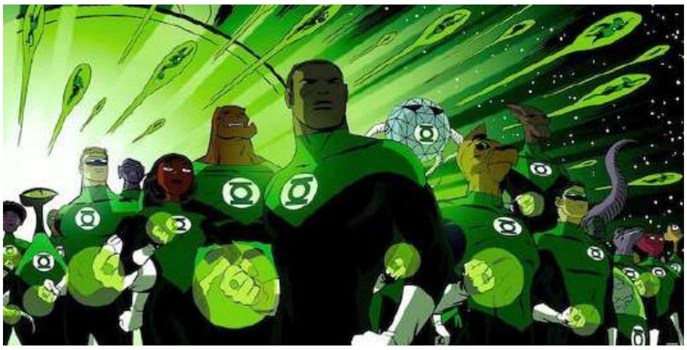


Figure 2.8: Diversity of characters [15]

Related Work

Character appearance can change between panels. Characters may be replaced by some other in next books. Characters can be non-humanoid with differing poses and colors making it one of the difficult element to detect. Sun et al. [102] used SIFT [103] to find characters. Qin et al. [104] used CNN based Fast R-CNN network [94] to find faces on Manga 109, but finds it difficult to generalize on Western comics due to diversity of characters. Chu et al. [105] used custom bounding boxes and combined traditional image features with CNN to improve upon Qin et al. [104]. Nguyen et al. [106] used Fast R-CNN based network [94] to learn multiple objectives on eBDTheque along with character which helped generalize a bit better. Even though CNN based techniques perform better than chance, they still fail to generalize cross domain or for slightly complex characters.

[15] `https://www.scifipulse.net/in-review-green-lantern-corps-37/`

2.8.5 Balloon

Balloons refers to enclosed areas that contains discussion between characters or emotions conveyed. They come in various shapes and sizes. Shapes are non-uniform across different books and comes in different shapes, sizes, colors. Artist can design comics without balloons as well, but balloons add another dimension of expression. They can contain text as well as onomatopoeia. Balloons can be broadly classified into speech balloons, thought balloons and narration as shown in fig 2.9. Speech balloons convey dialogue between characters on the panel, though bubble captures thought in character's mind and narration signifies narrator voice or scene change. Speech and though balloons point their towards it's character to tell user it's origin [107].

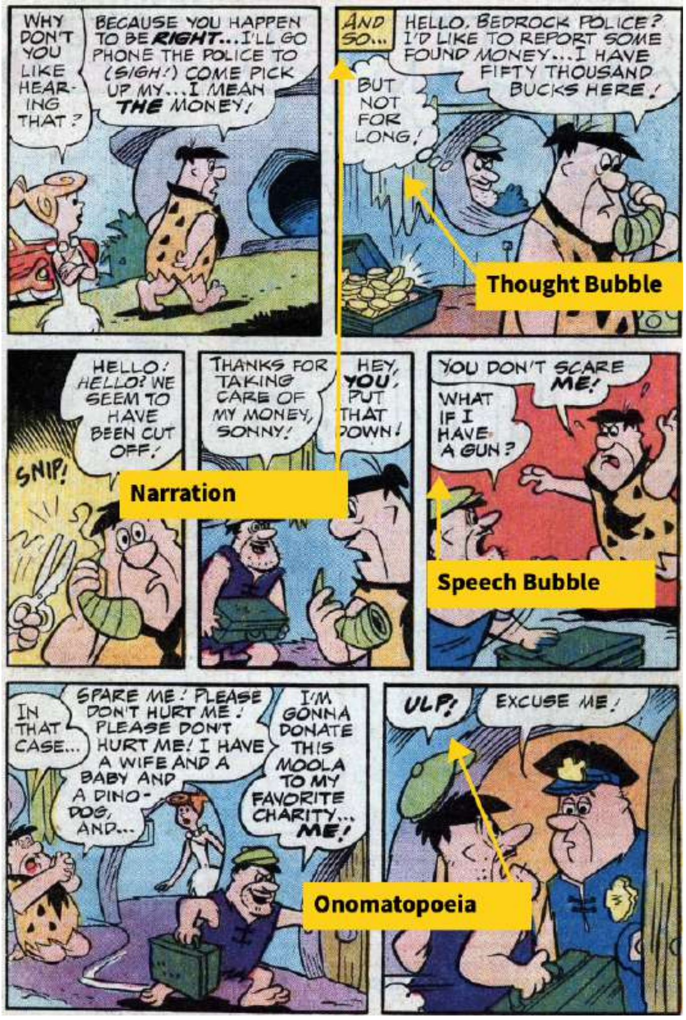

WHY DON'T YOU LIKE HEAR-ING THAT?
BECAUSE YOU HAPPEN TO BE RIGHT...I'LL GO PHONE THE POLICE TO (SIGH!) COME PICK UP MY...I MEAN THE MONEY!
AND SO...
HELLO, BEDROCK POLICE? I'D LIKE TO REPORT SOME FOUND MONEY...I HAVE FIFTY THOUSAND BUCKS HERE!
BUT NOT FOR LONG!
Thought Bubble
HELLO! HELLO? WE SEEM TO HAVE BEEN CUT OFF!
SNIP!
THANKS FOR TAKING CARE OF MY MONEY, SONNY!
HEY, YOU! PUT THAT DOWN!
Narration
YOU DON'T SCARE ME!
WHAT IF I HAVE A GUN?
Speech Bubble
IN THAT CASE...
SPARE ME! PLEASE DON'T HURT ME! PLEASE DON'T HURT ME! I HAVE A WIFE AND A BABY AND A DINO-DOG, AND...
I'M GONNA DONATE THIS MOOLA TO MY FAVORITE CHARITY... ME!
ULP!
EXCUSE ME!
Onomatopoeia

Figure 2.9: Speech Balloons in Comics[16]

Related Work

Balloons refers to enclosed areas that contains discussion between characters or emotions conveyed. They come in various shapes and sizes. Shapes are not standard and do not usually used across different books. Ho et al. [93] attempted to find speech balloons through heuristics of connected components. They used pixel ratios, size and shapes to locate and extract speech balloons and text for Manga. However, their approach could not be generalized to other comics. Correia et al. [108] used edge detection and histograms to locate speech balloons and associate them to the speaker. Rigaud et al. [109] used topological and spatial information of connected components and combined with adaptive threshold to find speech balloons independent of the text. However, this method occasionally produces false positives if the image contains centered alphabet looking elements. Iyyer et al. [11] used CNN based Fast-R CNN to extract speech balloons. Further they experiment with OCR Abby [17], Google cloudvision [18], Tessearct [19] to extract text from each panels. Google Cloud Vision OCR performs better than others, even though it still has lots of spell mistakes due to unique fonts and scan quality of comics. They apply heuristic based spell correction. Nguyen et al. [106] used CNN based Mask R-CNN [110] to extract multiple elements speech balloon, character, panels by learning primary task association with speaker on DCM772 and eBDtheque datasets. However shape of speech balloons are not captured. Dubray et al. [111] used UNet and VGG 16 [112] based encoder to segment each pixel as speech balloon for GNC dataset. This helps to recognize various shapes, but performs poorly on speech balloons with bright areas or whose color matches closely to background. Dutta et al. [113] combined edges, shapes from supplementary CNN with main UNet to improve performance of speech balloon detection. This helped to alleviate problems of Dubray et al. [111] on eBDtheque, DCM and Manga 109 datasets.

2.8.6 Temporal Flow in Comics

The idea of temporal flow is fundamental to how comics depict time passing and the sequential aspect of the story. Time is conveyed in comics through a variety of graphic and narrative strategies that help readers recognize and comprehend how events develop over time. Panel placement, transitions, and the usage of temporal

[16] https://flintstones.fandom.com/wiki/The_Flintstones_(Marvel_Comics)

[17] https://www.abbyy.com/ocr-sdk/

[18] https://cloud.google.com/vision

[19] https://github.com/tesseract-ocr/tesseract

indicators like speech, captions, and sound effects all serve to convey time as shown in fig 2.7. Too much blockage will make different readers follow different reading order [87]. Size, shape, and placement of panels can indicate the length of an action or the pace of a scene. Shifts in time and location are indicated through transitions, such as those from scene to scene or action to action [114]. The story is better anchored in a certain time period when there are temporal markers present, such as clocks, calendars, or visual clues like changing weather or attire [86].

Related Work

Narrative arc would be twisted, turned and slowed by pacing the story at appropriate places. Pacing helps to maintain the curiosity, tension in the story. On a macro level, they also determine whether the book could be read quicker or not. Mikkonnen et al. [86] surveys various techniques of representing time from artists. Artists usually manipulate panel shapes to dilate or contract time. Smaller and more panels usually used to expand time whereas splash panels to create the central action scene. Panels could be spaced differently with more gutter space indicating dilation of time. Panel shape can be changed at last panel of the page to make it a page turner. To confuse users, panels can also be not arranged in order and would be placed in a floating manner in the page. However there are no other works on extracting this as a feature from comics.

2.8.7 Onomatopoeia

Comic book creators employ onomatopoetic words frequently, a literary technique that mimics natural sounds, to enhance sensory details prevalent in visual storytelling among different scenes portrayed. Onomatopoeia's impact is enormous because it adds an additional audible dimension through descriptive words that mimic noises made from action scenes such as punches thrown “pow”, heavy impacts “bam” or footfalls “splat”. This becomes more important as comics are often translated between different languages, each culture has it's own way of expression. In other words, onomatopoetic words capture aesthetics of the culture and emotions portrayed by comic character [86]. As demonstrated on fig 2.4, using onomatopoeia to create audiovisual synergies during action-packed sequences engages and immerses the audience in the tale's fictional world, resulting in a memorable experience. Using onomatopoeic phrases captures the reader's engagement and attention beneath every action sequence [114].

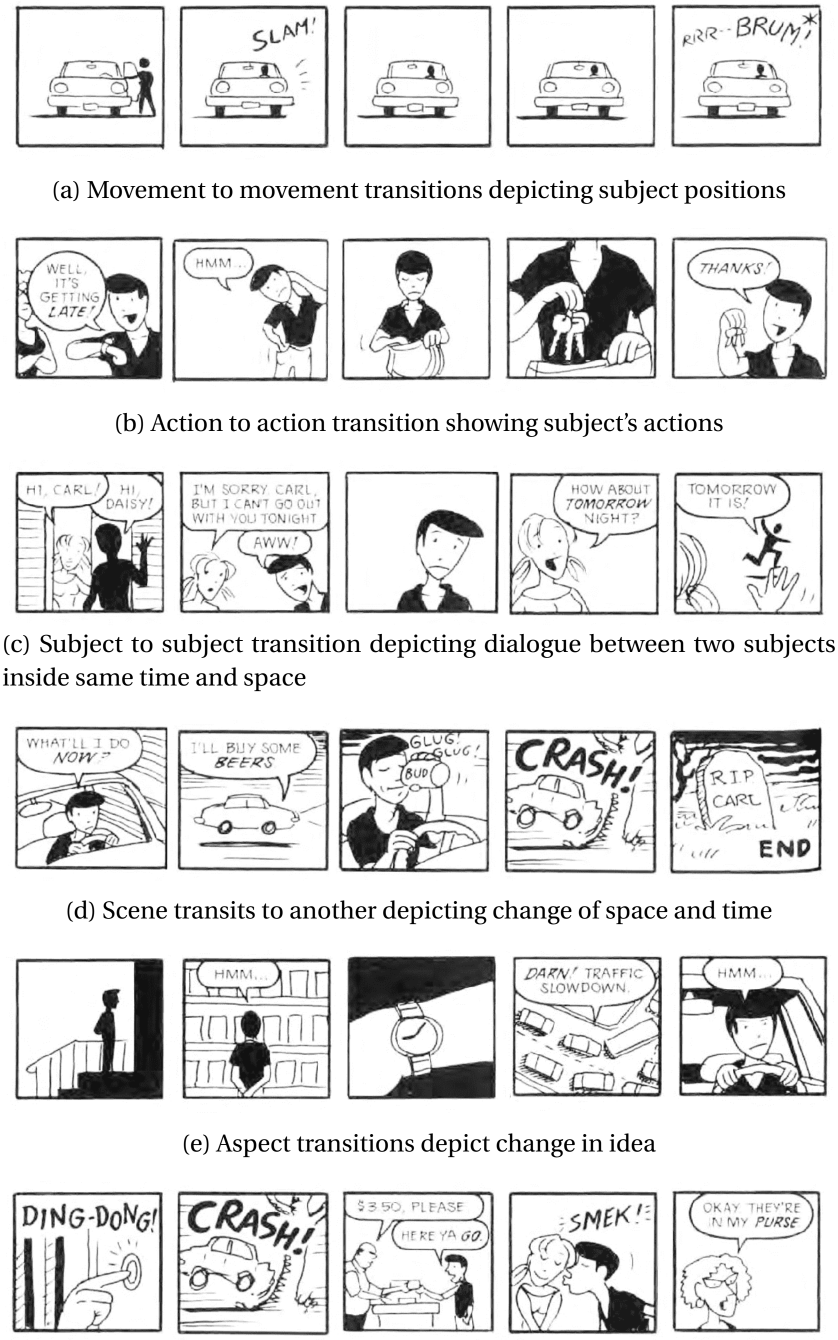
Related Work

Baek et al. [115] creates a dataset of manga consisting of onomatopoetic texts in curved, expanded, splattered, and deformed states. Later they propose a model to stitch related onomatopoetic texts together through link prediction and textual analysis using pointer networks [116]. Rohan et al. [117] attempts to assess impact of full onomatopoetic translation against partial textual translation on cognitive load and interest. They used eye-tracker and analyzed readers fixation and other eye-tracker metrics on comics with full translation and another with partial textual translation. Manga with full translation gathered more interest than the counterpart. Takizawa et al. [118] proposes a transformer based sequence-sequence model to convert human speech into onomatopoetic texts. This is helpful to create sound effects for media and other expressive presentations. Onomatopoetic expressions are also part of normal life in different cultures, Okada et al. [119] attempted to analyze onomatopoetic word emotions such as anger, sadness, peace of mind during different phases of covid-19. As second wave of covid began to phase out, anger and fear started to decrease with people using more expressions related to peace of mind. Such analysis are helpful to macro-analyze emotional trends on trending topics over time.

2.8.8 Transitions

Comic book readers can enjoy a more immersive and engaging story with the effective use of comic panel transitions. These transitions serve as important tools for directing the reader's comprehension and involvement in the narrative by conveying changes in time, action, and spatial connections across the panels. To capture fluidity of motion. Panels may utilize Movement to Movement transitions to depict several stages or positions of a subject or object. Alternatively. Action to Action transitions may be employed to create a dynamic impression of movement by highlighting continuous action or motion across panels. Subject to Subject transitions enable the story to provide different perspectives by moving from one subject to another. Meanwhile. Scene to Scene transitions transport readers across distinct settings and different moments in time. Advancing the narrative. In order to preserve character continuity. Continued Conversational transitions promote conversational consistency between frames. On the other hand Disconnected Transitions are intentional pauses or juxtapositions between frames that can elicit specific impacts or contrasts as shown in fig 2.10. Overall these varied transition techniques have an impact on narrative structure, atmosphere and tempo in comic book storytelling [114].

Figure 2.10: Illustration of different panel transitions from Mccloud et al. [20]



(a) Movement to movement transitions depicting subject positions

(b) Action to action transition showing subject's actions

(c) Subject to subject transition depicting dialogue between two subjects inside same time and space

(d) Scene transits to another depicting change of space and time

(e) Aspect transitions depict change in idea

(f) Non-seqitor corresponds to bunch of disconnected scenes

Related Work

Iyyer et al. [11] attempts to decipher narratives in comic book using both visual and textual features. First they create a dataset consisting of panels and its associated text. They annotate five transitions such as movement-movement, subject-subject, action-action, scene-scene and conversation in their dataset. They stitch both panels and text and provide closure to the disparate narrative. Three cloze style tasks are designed and evaluated. Text cloze detects appropriate dialogue for the current panel provided previous text and panel. Visual cloze task is designed to detect panel image provided its previous panel images and their associated text. Finally, they design a task to reorder and correctly place dialogue for each character in the panel. They achieve better than random chance results in all three tasks, albeit could not pass human performance. Devi et al. [120] generates unsupervised abstractive dialogues that represent the entire comic story while maintaining its essence. Individual elements from comics such as panels, text, speech balloons are first extracted. Later they score each word in a sentence and form a graph with edges representing it's score. It focuses on removing redundancy, ensuring informativeness, and using appropriate edge weights and path re-ranking methods. The goal is to create concise summaries that capture the essential information and relationships between words. This approach differs from previous supervised methods and incorporates a language model to improve the fluency of the generated sentences.

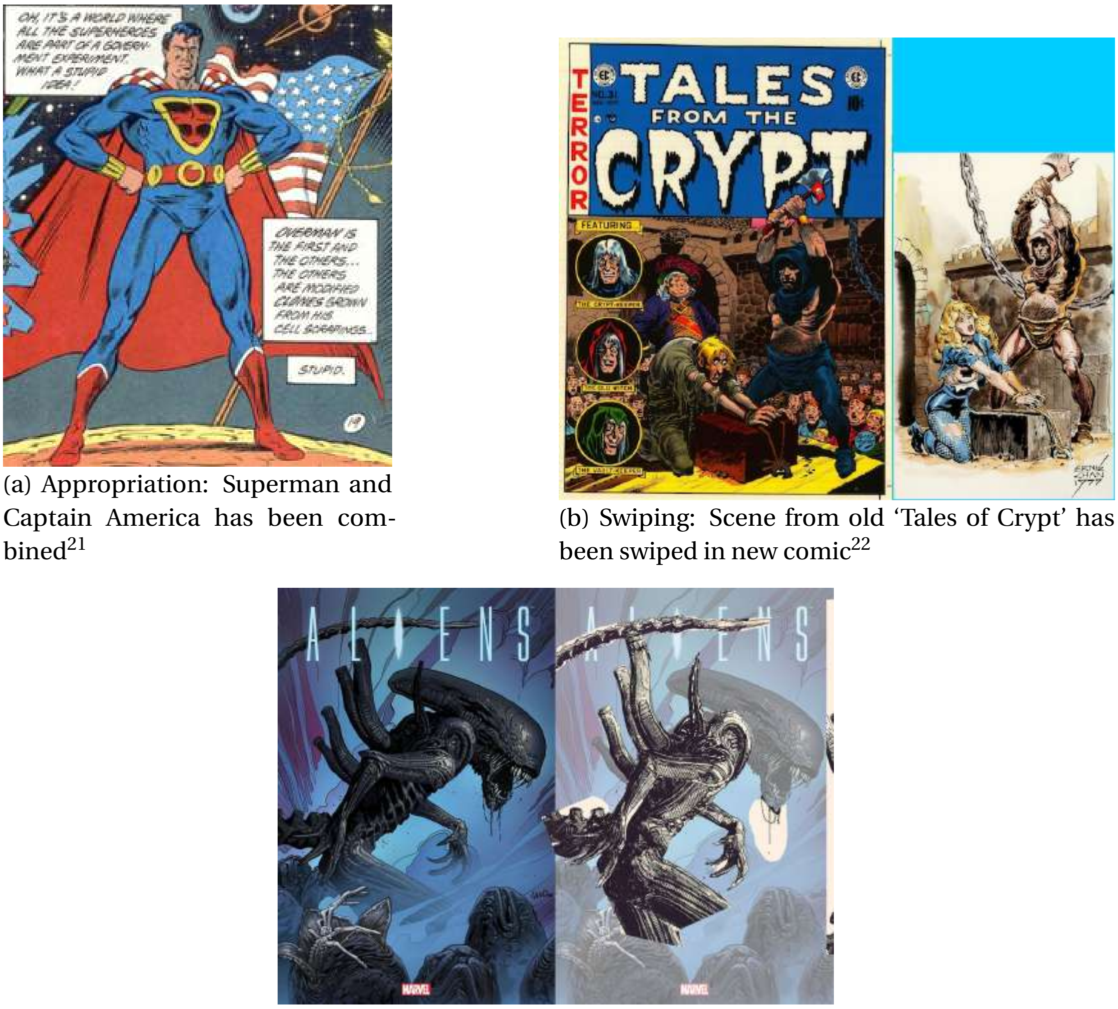
2.8.9 Artistic Style

Artistic style (Stylometery) is analysis of linguistic or visual styles across different documents. Study of writing style, authorship attribution based on textual and visual features, plagiarism detection and genre detection are some of the areas of it's application. Textual features such as word frequency, sentence length, and syntactic structures are analyzed to represent textual stylometry [121]. Visual stylometery initially focused on analyzing fine arts such as "Van Gogh" paintings through texture, geometry, style, color and other attributes [122]. This is really important in comics industry as art is usually plagiarized from other artists. Typically, easiest form of plagiarizing is Cloning, wherein artist tend to switch characters, but do not change their pose or semantic meaning. Appropriation is also another form of plagiarizing, where different characters are merged to form another unique character. In a famous saga, marvel comics plagia-

[20] https://scottmccloud.com/2-print/1-uc/

rized famous artist 'Tristan Jones' Alien comics through photo-shop editing. Different forms of plagiarizing are showed in below fig 2.11

Figure 2.11: Different Forms of Plagiarization


OH, IT'S A WORLD WHERE ALL THE SUPERHEROES ARE PART OF A GOVERNMENT EXPERIMENT. WHAT A STUPID IDEA!
OVERMAN IS THE FIRST AND THE OTHERS... THE OTHERS ARE MODIFIED CLONES GROWN FROM HIS CELL SCRAPINGS.
STUPID.
19
TERROR
TALES FROM THE CRYPT
FEATURING
ALIENS
ALIENS
MARVEL
MARVEL

(a) Appropriation: Superman and Captain America has been combined[21]

(b) Swiping: Scene from old 'Tales of Crypt' has been swiped in new comic[22]

(c) Cloning: Alien artwork cloned from Tristan Jones by Marvel[23]

[21] Appropriation Source: https://www.plagiarismchecker.net/dev/plagiarism-and-swipe-comics/

[22] Swiping Source: https://boingboing.net/2020/04/30/heres-another-example-of-one.html

[23] Cloning Source: https://aiptcomics.com/2020/09/08/tristan-jones-interview-greg-land/

Related Work

Sun et al. [123] attempted to find duplicated anime styles in books through HOG [124] feature similarity. Dunst et al. [125] tried multiple task of discerning discern genre and artist from classical features of comics through shapes, edge histograms, color layout and other classical features. Laubrock et al. [126] used combination of CNN based inception net V3 [127] and Fully connected neural network to learn on multiple tasks such as finding author, illustrator and genre on GNC corpus. They were able to get better results by using features from middle layers of inception net rather than last layer, thereby demonstrating that last layer mostly captures higher level semantic details whereas middle layers tend to focus on textures, colors and other details. Dunst et al. [128] combined textual features such as count of words per page, length of words and type to token ratio along with previous traditional image features to improve performance on authorship classification drastically. However, genre classification performance remained the same.

2.8.10 Book Cover

Attractive book cover for comics acts as lightning rod for readers. Readers often tend to pick up comic books from display based on its cover. After direct market selling started by comics, book covers became even more important. Comic book publishers specially created book covers for their own outlet. Furthermore, a good book cover will create purchase impulse and keep user guessing about story line. Book covers work on principles of contrast, focus and emotion. They contrast out non-important characters from story and focus on most important part of story-line. Finally they keep readers on tenterhooks through emotion [24]

Related Work

Yang et al. [129] extracts book cover textual information through threshold and segmentation. Furthermore, they apply OCR to extract text and classify the book. Iwana et al. [130] collected book covers and their categories from amazon.com [25]. They tried to classify the book into its categories using CNN based Alexnet [131] with above chance performance. Color, Objects, Textual elements from book cover were shown to impact the models performance more. Biradar et al. [132] combined textual title along with

[24] https://comicalopinions.com/what-makes-a-good-comic-book-cover/

[25] https://www.amazon.com/

CNN based features extracted from book cover image to improve models classification performance.

2.8.11 Genre

Genres are classifications that encompass domain and its associated writing style, established with rules and conventions over time. Common genres in comics include superhero, detective, action, adventure, romance, fantasy, children, humor and many more. Two books are considered similar, most likely they belong to the same genre. Therefore, accurately identifying the genre is crucial in determining book similarity. Comic books could be categorized under multiple sub-genres explicitly or implicitly. Some books contains multiple short stories belonging to different genres, others have narrative that may reflect multiple genres over time [133].

Related Work

Genre is one of the important facet used to choose the document. Sivaraman et al. [134] used combination of multiple CNN based model features to predict genre from movie posters from Youtube trailer dataset. Bucher et al. [135] created a domain specific features such as parts-of-speech, adjectives, pronouns, sentence length, casing and other lexical features to classify genre with improved performance over Gutenberg corpus [26]. Barney et al. [136] used movie posters from Movielens dataset [137] to classify genre with improved performance. Xu et al. [138] treats comics as sequence of pages which in turn a sequence of texts. First they create a dataset scraped from internet. They also clean comic books and remove books with partial pages. Genre are manually categorized into few important ones. They use sequence information and cascade to book level using multiple CNN based Resnet50 [139] models. Finally they use weighted voting to determine genre of the comic book with an improved F1-score. Dey et al. [140] used word embeddings and averaged them for every chunk to represent as genre for fiction books from Gutenberg corpus.

2.8.12 Gender

Readers often have preferences when it comes to the gender of the protagonist or the overall focus of a story, whether it's centered around males, females, or third gender/children. For example: Wonder Woman, Blondie, and Sheena feature powerful

[26] `https://www.projekt-gutenberg.org/`

female characters who play a central role in driving the narrative, whereas characters like Captain Savage and Batman have different dynamics. Additionally, children may have specific interests in characters from Disney and other franchises. Gender becomes a significant aspect for users when searching for content [141].

Related Work

kondreddi et al. [142] attempts to smoothen process of crowd sourcing information from humans using a combination of information extraction and Higgins engine. They create a dataset about movies by scraping Wikipedia. Personal pronouns and relationships are extracted. A relationship triple is formed and later used to validate triples. Jockers et al. [143] extracts male and female pronouns along with niche action verbs to classify genre. Wu et al. [144] study analyzed characters in Chinese martial arts novels to achieve automatic Q&A with fictional characters. Using a corpus of 1,435 novels and character vectors generated through Skip-gram model training, the research explored similarity among characters from the same author and successfully predicted gender using logistic regression and support vector machine algorithms. Xu et al. [145] built open-access toolkit that automatically extracts and aligns narrative events in chronological order. It provides visualizations and facilitates navigation between graphics and text, making it useful for understanding narratives, question answering, and bias analysis. The toolkit has been validated through human evaluations and offers both a python library and a user-friendly web interface.

2.8.13 Super Sense

Super sense refers to assignment of coarse-grained semantic labels or themes to words in a text, enabling a macro understanding of language. It involves identifying the specific sense or meaning of words beyond their basic lexical categories [146]. Super sense tagging has various applications, including question-answering, information retrieval, and improving machine translation. It aids in disambiguation of homonyms and resolving syntactic ambiguities. For instance, it can differentiate between the noun financial “bank” and the verb parking “bank” [147]. A list of noun super senses are shown in table 2.2 and verb related super sense are shown in table 2.3 from Ciaramita et al. [148]. Adjectives and Adverbs are also captured as a category, these would be incredibly useful to retrieve books that may use more adjectives. For example: fantasy books or Humor books

Table 2.2: Supersense categories related to nouns [148]

Category	**Related nouns**
act	acts or actions
animal	animals
artifact	man-made objects
attribute	attributes of people and objects
body	body parts
cognition	cognitive processes and contents
communication	communicative processes and contents
event	natural events
feeling	feelings and emotions
food	foods and drinks
group	groupings of people or objects
location	spatial position
motive	goals
object	natural objects (not man-made)
quantity	quantities and units of measure
phenomenon	natural phenomena
plant	plants
possession	possession and transfer of possession
process	natural processes
person	people
relation	relations between people or things or ideas
shape	two and three dimensional shapes
state	stable states of affairs
substance	substances
time	time and temporal relations
Tops	abstract terms for unique beginners

Related Work

In Computational Literary Studies, theme analysis is frequently done through topic modelling. However, this method is often used for exploratory purposes rather than hypothesis testing. Holloway et al. [149] presents a comprehensive framework that promotes and explains ‘aboutness’ in narratives. Through empirical validation of a new Corpus of German Novellas, this study aims to demonstrate the usefulness of topic modelling in analysing ‘aboutness’ in fictional texts. Rizoiu et al. [150] addresses issues in terms and synonyms layers by investigating subject extraction for ontology learning and presenting available methods. It suggests an integrated approach as the first step

Table 2.3: Super sense categories related to verbs [148]

Category	Related verbs
body	grooming, dressing and bodily care
change	size, temperature change, intensifying
cognition	thinking, judging, analyzing, doubting
communication	telling, asking, ordering, singing
competition	fighting, athletic activities
consumption	eating and drinking
contact	touching, hitting, tying, digging
creation	sewing, baking, painting, performing
emotion	feeling
motion	walking, flying, swimming
perception	seeing, hearing, feeling
possession	buying, selling, owning
social	political and social activities and events
stative	being, having, spatial relations
contact	touching, hitting, tying, digging
weather	raining, snowing, thawing, thundering

towards idea development and ontology learning for extracting significant concepts. In order to bridge the conceptual and topical divide, future research directions are explored.

Badawy et al. [151] extracts key themes from textual learning materials while also connecting relevant resources to create dynamic knowledge graphs with higher accuracy than current techniques resulting in self-learning and long-term community growth promotion. Jockers et al. [152] analyzes a corpus of 19th-century novels in search for themes that could represent the novels using LDA [153] to extract topics from books.

2.9 Summary

We elaborated on current state of the art research on comic book analysis, information retrieval and explainability in previous sections. We start with “How do one represent a Comic Book?”. Defining facets that comprise a comic book could help represent it. Researchers focus on panel and speech balloons extraction as they are basic semantic building blocks. Low level features such as color, texture and edges are used to find artistic style of the book along with the text. Characters detection in comics is difficult due to deformations of character and limited dataset. To determine narrative, transitions like scene-scene or action-action are predicted with limited success. Textual elements such as shape of balloons, dialogue text are often used to predict genre. On the contrary, narrative, page layout and temporal elements are seldom used. Textual elements are not fully utilized, that could represent other facets such as gender, super sense and topics [154].

Users have diverse preferences, they may seek specific facts or want to explore. Lookup based search is usually used to search facts. Interactive retrieval employed to iteratively refine their goals and communicate context, and exploratory searches to enhance information discovery.

Marking documents as relevant or irrelevant explicitly to provide feedback could lead to fatigue. Comics contain both images and text, hence a single mode of feedback is insufficient [155]. Explicitly marking documents could be intrusive, leading to fatigue. Implicit signals from screen activity are often used as feedback, although explicit feedback is more direct [156]. Limited data availability for training and evaluating feedback models hampers their effectiveness and precision. Feedback alone may not suffice as the system should be adaptable. Constraints during runtime may require simplicity, and achieved through GD. List-based user interface presents search results in a ranked list, which can bias users towards the top results. Graph-based UI allows users to explore relationships between documents and avoids ranking bias. Map-based UI enables exploration on a map, but may be affected by filter bubbles.

Explainable AI (XAI) approaches vary based on the scenario, such as using SHAP, LIME, and Feature Importance for classification explanations. In information retrieval, explaining rankings and relevance is challenging due to factors like context and user needs. Obtaining ground truth for explanations is challenging in XAI. System presents a few relevant books to user, but *"Why does system believe these books are relevant to the query?"* is an open question. Researchers attempt to answer this through various modes of explanations. Explaining personalization is largely unexplored. Users may have some thoughts on *"What do the system think about my likes and dislikes?"*. Opaque personalization due to feedback could also cause dissatisfaction among users, due to loss of transparency [9].

Finally, we also discuss on user study and concentrate on key measures such as contextual which describes demographic and background information. Interaction based measures such as dwell time, click rate and others signify engagement. Usability metric consists of qualitative metric like satisfaction and more quantitative metrics like effectiveness and efficiency that can measure overall performance of the system.

3
Methods, Materials and Models

We begin with existing dataset and create our dataset for retrieval. We propose our model of interactive information retrieval on comics with explanations. We elaborate on facets extraction approach. Later, we discuss its implementation details carried out fulfill research goals.

3.1 Datasets

We discuss on available datasets compiled for various comic book research from Augereau et al. [157] below.

3.1.1 eBDthèque

The eBDthèque dataset comprises 100 pages from 25 French, American, and Japanese comics spanning from 1905 to 2012. It includes annotations for panels, text, balloons, reading path, balloon shape, text transcription, along with metadata on album title, artist, release date. The selection of comics focused on representing a range of styles, lending itself to be a better evaluation dataset. eBDthèque was created by L3I group at La Rochelle [158]

3.1.2 DCM COMICS

Iyyer et al. [11] downloaded 4000 golden age books from Digital Comic Museum [27]. They manually annotated small portion (500 pages, 1500 text area) of these books. They used Fast R-CNN to learn extracting panels and text boxes on the small corpus. Later they used these models to extract panels and textboxes from all 4000 books. The dataset size is massive, but due to its semi-automated nature, data quality of the corpus is less.

[27] `https://digitalcomicmuseum.com/`

3.1.3 VLRC and TINTIN

The Visual Language Research Corpus (VLRC) is a carefully curated collection of over 36,000 coded panels from over 300 comics from diverse places and eras. It provides an in-depth investigation of the visual linguistic patterns used in several genres. Panel framing, semantic relationships between panels, page layout, multi-modality, and other aspects of visual languages are all covered in the annotations. The corpus is a good for researching and comprehending the many structural elements used in comics from various cultural perspectives. [28]

TINTIN Project [29] seeks to investigate multicultural trends in visual linguistics utilized in comics and their relationship to spoken linguistics. It looks into how people's languages and reading histories affect their ability to understand comics. Building on the Visual Language Research Corpus, the project makes use of the Multimodal Annotation Software Tool (MAST) to produce a corpus of annotated comics from more than 60 different nations. They aim to study 1000 books, Currently have analysed 285 books with 42000 pages [159].

3.1.4 Manga109

Aizawa et al. [160] from University of Tokyo created the Manga109 dataset, which is made up of 109 manga books, totaling more than 21,000 pages. Panel bounding boxes, character descriptions, and text content are just a few of the specific annotations that are present. Due to its breadth and quality annotations, the dataset is useful for models to learn. However, its limited by manga artistic style elements. Hence cannot be applied for European or American comics.

3.1.5 GNC

Graphic Narrative Corpus (GNC) consists of famous and critical 270 titles of North American english language graphic novels. GNC was created by Dunst et al. [97]. It consists of annotated panel polygons, characters, speech balloons, and text transcriptions as comments. Recordings of readers' eye movements are also included in the collection. Studies on illustrator classification, semantic segmentation, and eye movement modelling have all made use of GNC.

[28] `https://www.visuallanguagelab.com/vlrc`

[29] `https://www.visuallanguagelab.com/tintin`

3.1.6 Our Dataset

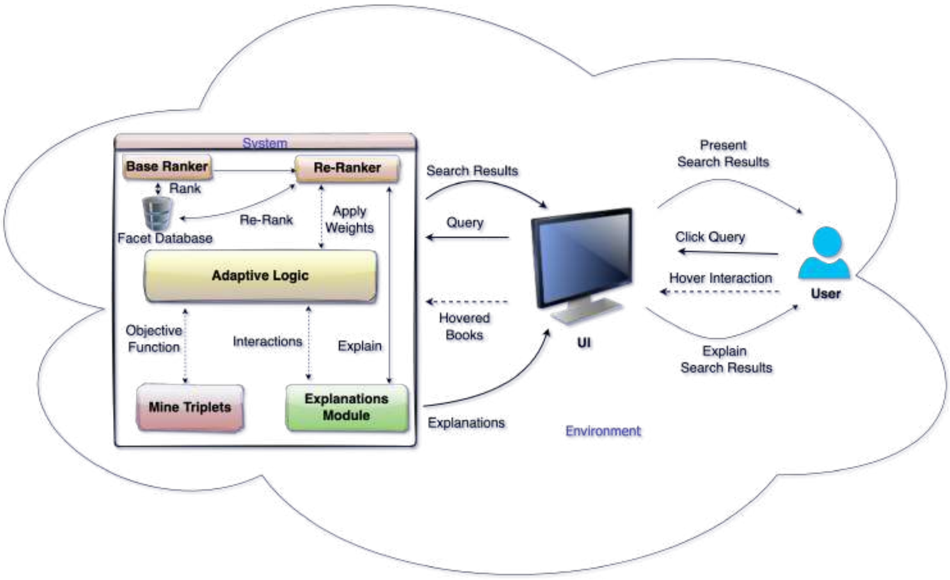
We use the 500 books from DCM COMICS dataset for our corpus. Since most of the books from DCM are not famous or known, We download a corpus of 1210 golden age and small number of silver age books from the internet. We apply techniques from Iyyer et al. [11] to extract panels, text from these books to maintain same format as DCM COMICS. These are famous books like Tin-Tin, Superman, Batman and others, that can be easily recognized even by non-readers or movie buffs. User can feel more familiar with the corpus, if there are some famous comic books in them. Evaluation of search engines can be extended to people from non computer science and movie buff background.

3.2 Concept

Comic books have a very diverse set of audience. Some people watch movie adaptations rather than reading. Avid readers have their unique tastes, some like visual art, some for the dialogues. Artists tend to focus on foreground, background, pacing and other storytelling aspects. Furthermore, Users may not have defined goals and it tend to evolve as they explore system. For the system to understand context behind users query, It must become adaptable to context. A System **S** is *adaptable* iff an external User **U** can directly control **S** outcome by changing some parameters of **S**. **S** need to explain its internal working to **U**. **S** can adapt to context, by understanding context behind their search. **S** is *context adaptable* iff **S** presents different output given same input query, due to change in context [8]. Furthermore, it highlights that these adaptations in behavior are not random and driven by towards finding context using a metric.

Figure 3.1 provides a general overview of the adaptive comic information retrieval system. System needs to check for context and model it with respect to users query and environment. Relevance feedback algorithm needs to absorb these interactions and change the facet weights of the retrieval system that would best fit the context. Retrieval system needs to fetch and rank the documents according to user preferences. System takes user query and produces relevant search results. At each turn, system attempts to learn context from their activity and adjust the facet weights accordingly. User can learn more about the system and results through provided explanations. To manually self-evaluate feature and retrieval quality, We create a subset of well-known 50 comic books, consisting of books from all genres and time, among our dataset (manual-evaluation dataset) . We issue sample 5 queries to test. We use F1-measure and try to find relevant books among top-5 search results. We then iterate to find best fit for independent variable.

Figure 3.1: Conceptual model of the system



3.3 Proposal

3.3.1 Facet Extraction

We propose creating our own dataset in similar lines to DCM COMICS dataset. Hybrid dataset is created with 500 comics from DCM COMICS dataset and 1200 comic books of famous characters from internet to bridge gap of unfamiliarity. We extract low level image features like color, texture and edges, as well as text to form base feature space. We use Iyyer et al. [11] approaches to extract panels and text dialogues from our dataset. Later, We extract visual facet (feature) like book cover, temporal facet like story pace and textual facets like genre, gender and broad topics to enrich our representation.

3.3.2 Interactive Information Retrieval

We propose creating a two-tiered feature space comprising of low-level features as base feature space and human interpretable domain based facets for re-ranking, similar to Polley et al. [73]. An online adaptation algorithm based off GD finds out users interest and uninterest through their mouse activity similar to Stober et al. [13]. Adaptation emphasizes and ignores facets that represents users interest and uninterest. This acts as a two-way switch, wherein users could explicitly communicate their preference to the system.

3.3.3 Explainability

We propose using textual topics from book cover, to explain impact of users previous search activity on personalization. A visual explanation comprising of facet weights, explains relevance of search results to query on a global level. Visual chart is used to locally explain story pace between query and book from search result. A textual comparison table is provided to compare multiple books from search results on its genre, story pace, characters appearance locally. Finally, topics extracted from comic are presented to user to provide more information about the book [140].

3.4 Facet Extraction

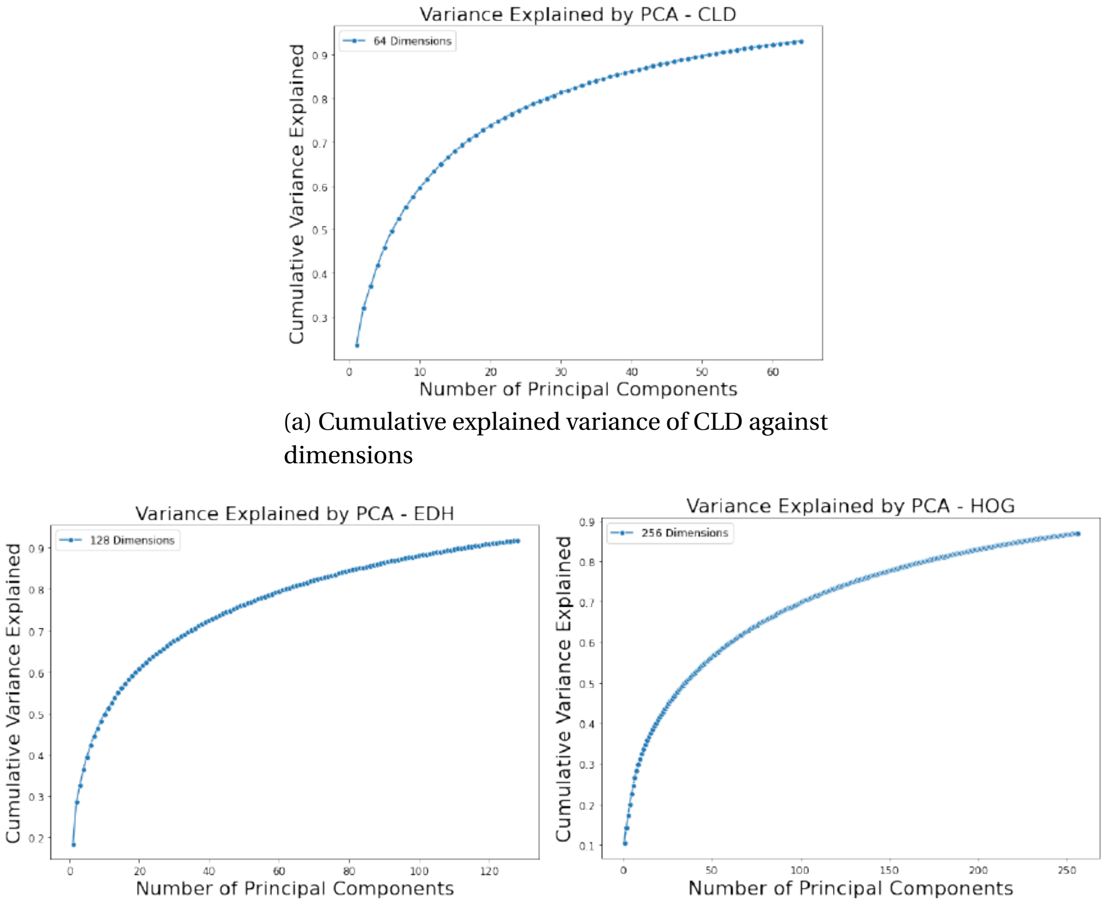
Our dataset closely resembles DCM COMICS and consists only 30 modern age comic books. We initially extract panel and text from each comic book using Fast R-CNN. We train model on annotated DCM COMICS dataset. We extract individual panel and its related text box image for each panel. Tessearct OCR [30] extracts text from each text box. Niche fonts and poor scan quality creates lot of spelling mistakes. We correct spelling mistakes manually by replacing single characters 1 by I, ! by I and 0 by o, non-english characters and numbers > 4 places are replaced with null. Multiple spaces and newlines are merged to single space and newline. FastPunct [31] and pyspellchecker [32] are used to correct remainder of spelling mistakes. Single consonant are merged with next word or deleted if next word is incorrect.

[30] https://github.com/tesseract-ocr/tesseract
[31] https://github.com/notAI-tech/fastPunct
[32] https://pypi.org/project/pyspellchecker/

3.4.1 Low Level Features

Visual Features

Figure 3.2: PCA explained variance plots for CLD, EDH and HOG against dimensions



(a) Cumulative explained variance of CLD against dimensions

(b) Cumulative explained variance of EDH against dimensions

(c) Cumulative explained variance of HOG against dimensions

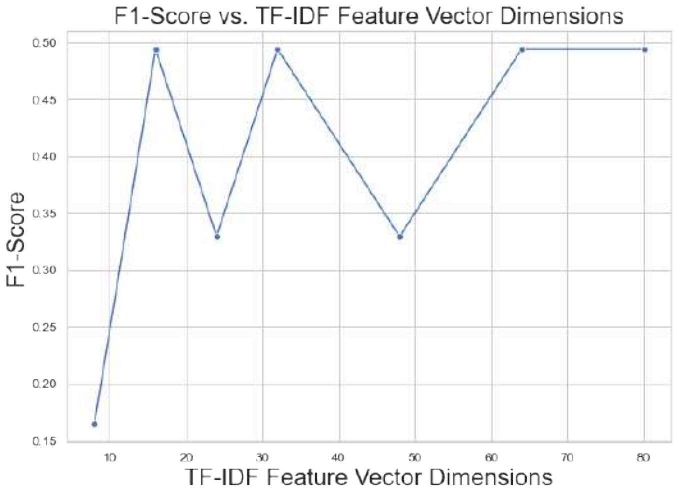
We implement of a base visual feature representation for comics using MPEG-7 standard. We extract various features such as Color Layout Descriptor (CLD) to describe color layout per panel [73], Edge Histogram Descriptor (EHD) to describe texture per panel [73] and Histogram of Oriented Gradients (HOG) [124] to describe coarse objects. We apply dimensionality reduction PCA to reduce each feature to CLD to 64, EDH to 128 and HOG to 256 dimensions based on percentage of variance plots 3.2.

We create a bag of visual words of 128 dimension each to get a fixed-length vector. We use weighted cosine similarity to compare all three features. Length and weights are set based on manual-evaluation dataset. We provide weights of 0.1 to CLD, 0.2 to EHD and 0.2 to HOG with intuition that EHD can represent style and HOG can represent objects.

Textual Features

We implement of a base textual feature representation for comics using TF-IDF feature vector. We concatenate all text from panels of single book into one. We also extract prompts describing topics using book cover using CLIP [25] based model from Huggingface [33]. Prompts from book cover could potentially bridge intent gap, as users may choose books based on its book cover. All text related to a book are combined with prompts. We apply stopword removal and tokenization using Scikit-Learn Library [34] and extract a 64 dimension TF-IDF vector [35] based on manual-evaluation dataset as shown in fig 3.3.

Figure 3.3: TF-IDF feature vector size influence on F1-score, We observe high and stable scores from 64th dimension feature vector size on wards



[33] https://huggingface.co/spaces/pharma/CLIP-Interrogator
[34] https://scikit-learn.org/stable/index.html
[35] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

3.4.2 Textual Facets

Genre

Comic book can have multiple genres. Some books are collection of multiple stories belonging to different genres. Categorization of genre is also not standard. Hence we first manually correct genre labels for all comics to have only *adventure, children, crime, detective, humor, jungle, mystery, non fiction, romance, sports, spy, superhero, vigilante, war, western* categories. We use Gensim Doc2Vec model and Logistic Regression [36] to train corpus on multi-class classification. We use default configuration for Doc2Vec model, albeit use `min_count=2` to remove extremely unique words. We stop training at xx epoch as per fig 3.4. We use reduce dimensionality of vector from Doc2Vec model to 16 dimensions as representation of genre 3.4. We select the model based on its retrieval performance on manual-evaluation dataset.

Figure 3.4: Details of genre facet creation

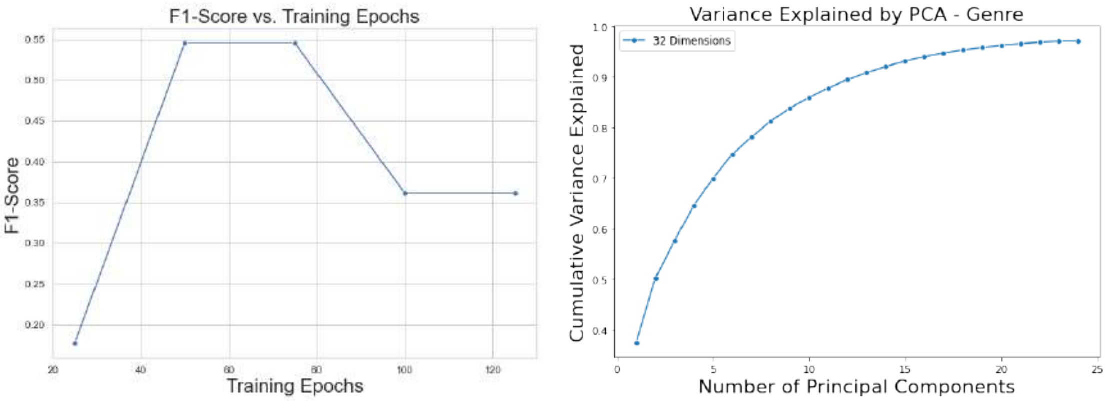


Figure 3.5: Training epochs influence on genre vis-à-vis F1-score

Figure 3.6: Cumulative explained variance cross 90% by 16 dimensions.

Gender

Comic books can have multiple humanoid and non-humanoid characters. Knowing them, would help to find relevant books. We use BookNLP (Bamman et al. 2014) [37] to extract referential gender for characters from book. BookNLP tags each character occurance with referential gender rather than exact gender. We categorize it further by

[36] `https://radimrehurek.com/gensim/models/doc2vec.html`

[37] `https://github.com/booknlp/booknlp`

summing up *("he", "him", "his") as Male, ("she", "her") as Female and ("xe", "xem", "xyr", "xir"), ("ze", "zem", "zir", "hir")* as Other gender and children. Counts are normalized for better representation in feature space.

Supersense

Supersense contain 26 labels for nouns, such as *animal, event or time* and 16 labels for verbs such as *competition, motion or cognition* and adjective. These labels are used to disambiguate each word in text and place them in one of the above coarse categories. Having such information about each book can give one a rough idea about topics discussed in the book. For example: Romance books may have more *emotion* related verbs and *relationships* noun categories, whereas action book may have more *competition* verb. We use BookNLP to find supersense of each word in the text. We count occurrence of super sense across each category and normalize them for better representation

Book Cover Prompt

Book cover prompts attempt to describe objects, colors and styles in the book cover in textual format. Since User could click a book based on its cover image, We can enrich this further by using textual representation of book cover. We use same book cover prompts extracted before. Prompt sentences are concatenated and encoded by a sentence transformer [161] with huggingface software [38]. We use a smaller model *all-MiniLM-L6-v2* to preserve computation to calculate sentence embeddings. We do not average multiple prompt embeddings into one due to loss of information. Instead, this feature space would be represented by collected of sentence-transformer embeddings through its concatenation.

3.4.3 Visual Facets

Book Cover

Book Covers are one of the primary source of attractiveness for a comic book. A good cover page along with splash panels are prerequisite for some comic fans. We use the feature vector from penultimate layer of the previously used, Clip model [25] *clip-*

[38] https://huggingface.co/

vit-large-patch14. Base model consists of a large Vision Transformer [24], We utilize penultimate layer of the model as facet to represent book cover.

3.4.4 Temporal Facets

Story Pace

Temporal element can be represented visually buy most of the comics. However, metadata about number of panels in each page, amount of textual information and number of pages can provide a good proxy to measure pace of the story. We create two features from this metadata information. One representing average panel per page per comic book 3.1 and another amount of panels and text in each book 3.2. These two features effectively represent story pace facet.

$$\mathbf{AP} = \frac{\mathbf{TP}}{\mathbf{TPB}} \tag{3.1}$$

Where **AP** is average number of panels per page in a book. **TP** is total number of panels in a book. **TPB** is total number of pages in a book.

$$\mathbf{TIPB} = \frac{1}{\mathbf{TP} + \mathbf{DTWB}} \tag{3.2}$$

Where, **TIPB** represents the total information per book, which measures the amount of information present in the panels and dialogue text of the book, **TP** represents the total number of panels in the book, **DTWB** represents the word count of the dialogue text in the book.

3.5 Interactive Information Retrieval

We first begin with defining Session and Turn. **Session**: Session represents a cohesive and continuous interaction between the user and the system, typically centered around a specific topic or task. User can close the session by closing browser or doing a direct book search from search bar. At session end, We refresh personalization to default setting. **Turn**: Turn refers to a single exchange or interaction within a session. It represents a pair of back and forth message between user and system.

At each turn, The context model attempts to learn users likes and dislikes in terms of facets represented by the system. Further it provides user preference to retrieval engine.

Retrieval engine is responsible for providing relevant comics that comply with user preference. At each turn, user interaction such as click, hover are tracked.

3.5.1 User Context Model

We utilize a multi-facet distance measure for comics. The involves constructing a complex measure by combining various facet, allowing for static computation of distances within each facet. User preferences are represented through weighting scheme of the aggregated facet distance. However, adjusting facet weights can be challenging for users, who may not have any final goal or know their own preferences. We define facet as a collection of individual features of a document. We attempt to formalize comic facet distance [13].

Definition

Let S be the feature space for set of documents D and X be the features. A facet x is a subset to S, $x \subseteq S$. Facet distance δ_x between documents $p, q, r \in D$ must satisfy following conditions.

1. **Positivity**: $\forall p, q \in D : \delta_x(p, q) > 0$

   $\forall p, q \in D : \delta_x(p, q) = 0 iff (p = q)$.

2. **Symmetry**: $\forall p, q \in D : \delta_x(p, q) = \delta_x(q, p)$

3. **Triangle Inequality**: $\forall p, q, r \in D : d(p, r) \leq d(p, q) + d(q, r)$

Facets X are captured with varied approaches. Even though they are normalized before, their distance aggregates to values that vastly differ from each other. For example: Facet distance $\delta_{bookcover}$ could be higher than $\delta_{storypace}$ due to its large feature representation. This can disproportionately bias results towards book cover. We normalize the δ_x using mean of facet facet distance as per 3.3 [13].

$$\delta'_x = \frac{\delta_x}{\mu_x} \tag{3.3}$$

Average facet distance can be pre-calculated by finding distance between every pair combination of documents from D as per equation 3.4. Extreme facet distances are ignored while creating the average to reduce bias.

$$\mu_x = \frac{1}{|\{(p,q) \in D^2\}|} \sum_{(p,q)\in D^2} \delta_x(p,q) \tag{3.4}$$

The distance between documents p and q in the set D for k facets $x_1, x_2, ..., x_k$ is determined by calculating the weighted sum of facet distances $\Delta_1, \Delta_2, ..., \Delta_k$ according to Equation 3.5.

$$\delta(p,q) = \sum_{i=1}^{k} w_i \delta_{x_i}(p,q) \tag{3.5}$$

Adaptation Goal

As user interacts with the system, they may click or hover on books they like, they may not pay attention towards books that they do not consider relevant. We hypothesize that user is interested in facet types exhibited by books they interacted with and not interested in certain facets exhibited by other books. Hence our model must make sure that query q and interested book a must be closer to each other than disinterested book b. Hence for a triplet pair $s(q,a,b)$, distance $\delta(q,a) \le \delta(q,b)$ [13]. Our goal is to generate facet weights such that, relative distance from query to interested book is nearer than for disinterested book. We can utilize a triplet loss function that can optimize relative distance towards the goal as per 3.6.

$$\mathcal{L}_{\text{triplet}}(q,a,b) = \max\big(\delta(q,a) - \delta(q,b) + m, 0\big) \tag{3.6}$$

where q is the anchor, a is interested book, b is disinterested book, m is the margin hyperparameter and δ is the distance metric.

We do not know explicit likes and dislikes provided by the user, hence do not have ready triplet pair. Schroff et al. [162] discusses below strategies to mine triplets.

1. **Random Sampling**: The dataset's anchor, interested book, and disinterested book samples can be chosen at random in order to mine triplets. Although this method offers a wide variety of triplets, it could also produce unbalanced or unhelpful samples.

2. **Hard Negative Mining**: The goal of hard negative mining is to locate difficult disinterested books that are more similar to the query than the interested book. It entails choosing the hardest disinterested book iteratively according to how close or far they are to the anchor.

3. **Semi-Hard Negative Mining**: Using a semi-hard negative mining strategy, disinterested books are chosen that are less close to the query than the interested book but still posing some challenge to the network. It makes sure that the chosen disinterested books are not either too easy or too difficult for training to be effective.

Assumption on User Interaction

We hypothesize, Any book user has examined during current session becomes their interested book and vice-versa. Due to mining from top-k result set, naturally most of the pairs comply with semi-hard mining as they were relevant to the query. We use a combination of semi-hard negative Mining when mild-disinterested books are available or else we sample at random.

Optimization with Gradient Descent

We can learn the facet weights through GD approach in the similar lines of [13]. Our objective function is as follows 3.7

$$\text{objective}(q,a,b) = \sum_{i=1}^{l} w_i(\delta f_i(q,b) - \delta f_i(q,a)) \tag{3.7}$$

In this formula, obj(s, a, b) represents the objective function with respect to the variables s, a, and b. The formula calculates the objective value by summing over l different terms, where each term is given by the product of the weight w_i and the difference between the function values $\delta f_i(q,b)$ and $\delta f_i(q,a)$ for a particular index i.

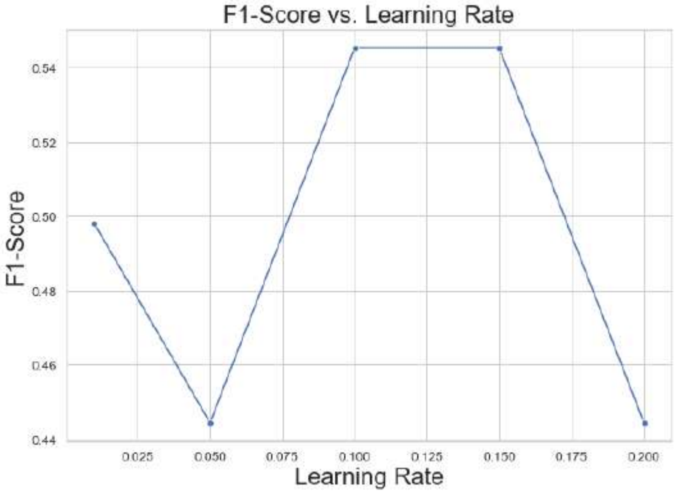
During learning phase, all mined triplet pairs are presented to GD algorithm until maximum epochs or convergence is reached. It tries to minimize relative distances by margin iteratively using update rule 3.8.

$$w_i = w_i + \eta \Delta w_i \tag{3.8}$$

$$\delta w_i = \frac{\partial \text{obj}(q,a,b)}{\partial w_i} = \delta f_i(q,b) - \delta f_i(q,a) \tag{3.9}$$

The learning rate η determines the size of each iteration step. To ensure the limits on w_i as specified in Equation 3.8 and Equation 3.9, an extra step is required after the update. This step involves setting all negative weights $\leq m$ to zero and then normalizing the weights, resulting in a constant weight sum of 1. We fix epochs as *1000*, as it takes lesser response time. We experiment with different learning rates on manual-evaluation-dataset 3.2 and decide that η of 0.1 is the better as shown in 3.7.

Figure 3.7: Gradient Descent learning rate's impact of output search results. Learning rate of 0.1 seems to work well



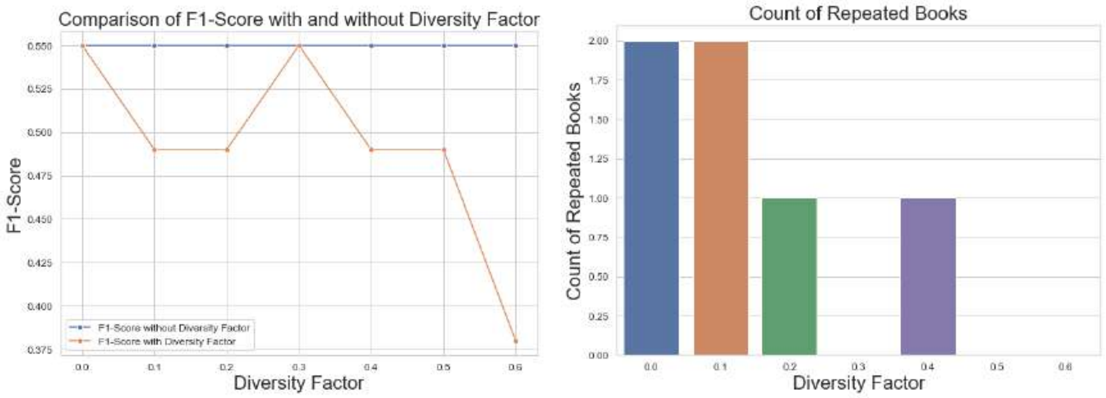
Algorithm 1 Optimization Algorithm

```
Require: $q$ = Query Book, Books hovered on by user : $D = h_1, h_2, ..., h_d$ treated as books of interest, Search Results
    from Low-Level Facet Ranker : $B = b_1, b_2, ..., b_m$
Let $TRIPLETLOSS$ refer to triplet loss function.
Let $VECTORIZE$ be a function that vectorize comic book into $k$ facet group.
Let $UPDATEWEIGHTS$ be a function that updates weights based on triplet pair in $h$ $TRIPLET\_PAIR\_LIST$.
Let $FACETWEIGHTS$ be updated facet weights $\mathbf{w}'_1, \mathbf{w}'_2, ..., \mathbf{w}'_k$ from $UPDATEWEIGHTS$.
Initialize individual_loss to 0 and total_loss to 0.
for $i \leftarrow 1$ to 1000 epochs do
    for $j \leftarrow 1$ to $h$ TRIPLET_PAIR_LIST do
        $q, a, b$ = TRIPLET_PAIR_LIST[$j$]
        $qiv, aiv, biv = VECTORIZE(q, a, b)$
        individual_loss = $TRIPLETLOSS(qiv, aiv, biv)$

        total_loss+ = individual_loss
        individual_loss = 0
        for $z \leftarrow 1$ to $k$ facets do
            $\mathbf{w}'_z = \mathbf{w}_z + \eta *$ total_loss
        end for
    end for
end for
Normalize weights $\mathbf{w}' = \frac{\mathbf{w}'}{\|\mathbf{w}'\|}$
Return normalized $FACETWEIGHTS$ $\mathbf{w}'$
```

Optimization algorithm for updating facet weights can be referred to algorithm 1 . It does not guarantee that constraints satisfaction after training nor guarantee global minima, as they have problems of getting stuck in local minima. On contrary, using slower learning rates may result in gradual change to weighting scheme, hence provide veneer of continuity in search results. Filter bubble is another problem of incremental update. Users can repeatedly see same books that they have seen before.

Figure 3.8: Impact of increasing diversity factor on repeated books. We see that F1-Score goes down if we try to decrease more. At 0.3, the F1-Score with and without diversity factor are the same albeit reduction in repeated books



(a) Comparison of F1-Score across different diversity factors against system without diversity factor

(b) Count of repeated books across different diversity factor

To reduce filter bubble, We add a diversity factor heuristically through running evaluation on manual-dataset 3.2 as shown in 3.8. We maintain a list of books that appeared in search results for three turns. If such book appears as search result in current turn, we reduce similarity score of the book by a factor set heuristically. We observe that a reduction in similarity of 0.3 before re-ranking removes repeated books and keeps the F1-score similar to its previous value. Reducing similarity score will decrease F1-score. Although, reducing similarity score artificially may not solve the issue completely.

3.5.2 Search Engine Implementation

We implement the retriever as shown below. Implementation is illustrated in fig 3.9, code could also be found at mentioned GitHub repository[42]

[42] https://github.com/surajsrivathsa/thesis_deployment

1. User hovers on initial search results to find more information. They may select a book as query or directly select a book from search bar.
2. Frontend system will cache all the hover and clicking activity and passes it on to backend server.
3. Base Ranker from Ranker module utilizes low-level features to select top-200 books for the query and sends it to re-ranker module.
4. Adapter sub-module mines triplet pairs from previous hovered bookx and optimizes new facet weights.
5. Re-Ranker sub-module calculates similarity score for top-200 books, reranks using new facet weights and adjusts with diversity factor.
6. Top-14 books are returned back from ranker module.
7. Explanations module uses interactions and top-14 books and returns explanations.
8. Backend server returns both top-14 results along with explanations to frontend.
9. User continues their search with new result set.

Step 1: Interaction with System

We track users mouse activity such as Click and Hover. If a user hovers on a book, then it is considered as interested book, else disinterested. This is core to our system. At each turn system tries to learn preferences. User interactions are not stored in any database due to GDPR. We implement the backend system using Python [163] based library FastAPI [39]. Frontend user interface and tracking system is built using ReactJS [40]. We containerize the application using Docker [41] to deploy on the server.

Step 2: Retrieve with Base-Ranker

We use low-level features such as visual bag of words made up of color, texture, histogram of gradients and textual bag of words made up of dialogues to retrieve top *200* books from corpus. *Cosine similarity* function is used to pick relevant books. We

[39] `https://fastapi.tiangolo.com/lo/`

[40] `https://react.dev/`

[41] `https://www.docker.com/`

filter base-ranker results to 200 to reduce burden on semantic re-ranker and noise for adaptation algorithm as referred by base-ranker algorithm 2.

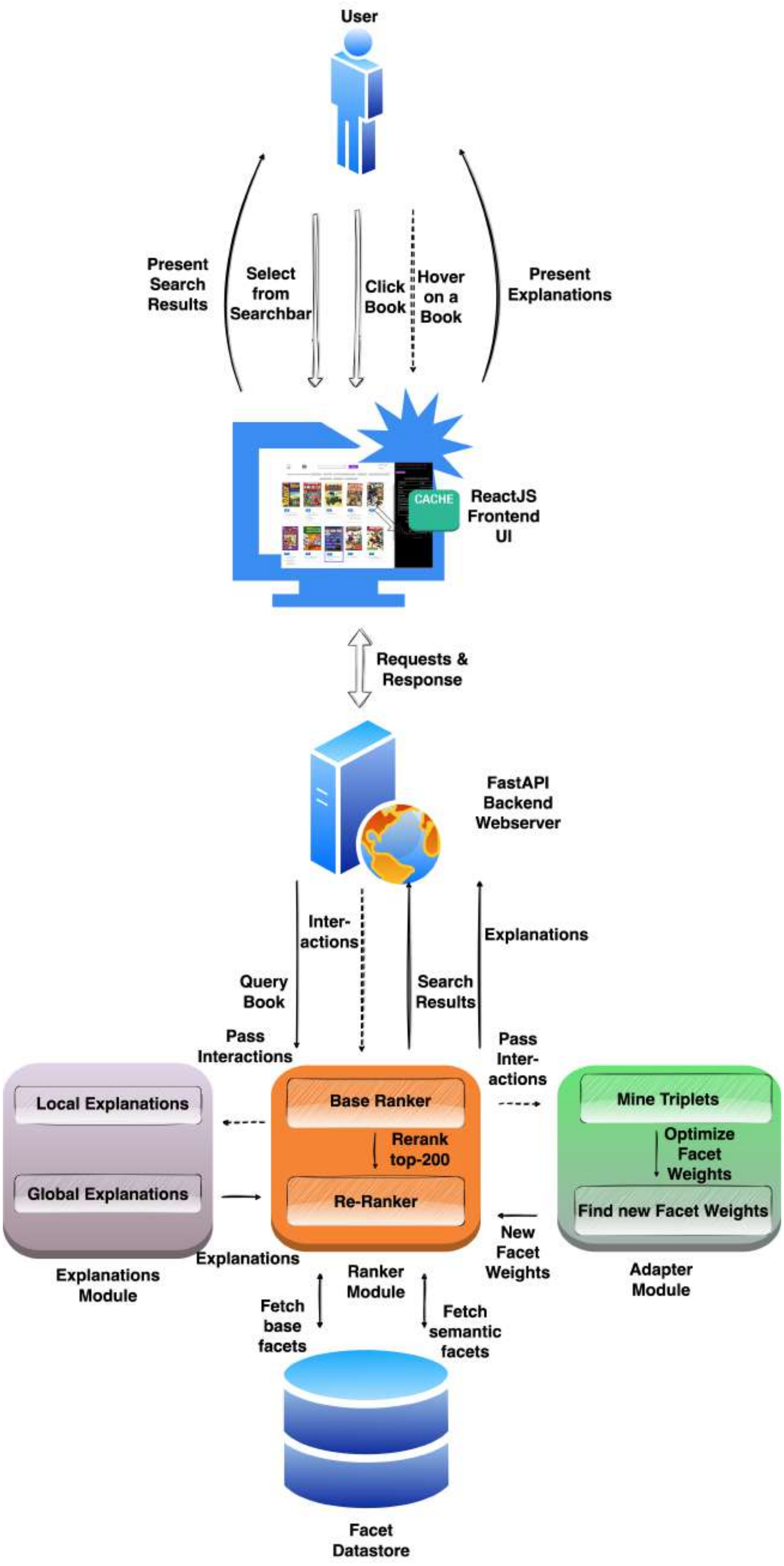
Step 3: Adapt and Re-Rank with Semantic Re-Ranker

Initial Results from Base-Ranker are reduced to top *200* books to minimize time and effort required by Re-Ranker. We mine triplets and adapt facet weights using optimization algorithm 1. Weights reflect interests of user for current turn. This is used to calculate weighted similarity between query and top *200* books. Results are re-ranked again and only top *14* books are presented to user as explained by re-ranking algorithm 3. Facet information are summarized in table 3.1.

Algorithm 2 Base Ranking Algorithm

```
Require: $q$ = Query Book $q$, Books hovered on by user : $D = h_1, h_2, ..., h_d$ treated as books of interest, book from
    corpus : $B = b_1, b_2, ..., b_z$ where z is total books in corpus
Ensure: Result after Base-Ranking $b'_1, b'_2, ..., b'_m \in B'$ where m = 200
Let $BASEFACETWEIGHTS$ be constant $FACETWEIGHTS$ set heuristically beforehand.
Let $VECTORIZE$ be a function that vectorize comic book into $k$ facet group.
Let $HistoricalSearchResults$ be list of search result books from past three turns.
$qv = VECTORIZE(q)$
for $i \leftarrow 1$ to $z$ books do
    $biv = VECTORIZE(bi)$
    Initialize facet_similarity score to 0, total_book_similarity to 0 and total_similarity_scores_list to [].
    for $j \leftarrow 1$ to $d$ facets do
        facet_similarity = cosine($qv, biv$) $*$ $BASEFACETWEIGHTS[j]$
        if $bi \in$ HistoricalSearchResults then
            facet_similarity = facet_similarity − 0.1
        end if
        total_book_similarity = total_book_similarity + facet_similarity
        Append total_book_similarity to total_similarity_scores_list
    end for
end for
Return top 200 $B'$ by sorting based on total_similarity_scores_list
```

Figure 3.9: AESC system design



Algorithm 3 Re-Ranking Algorithm

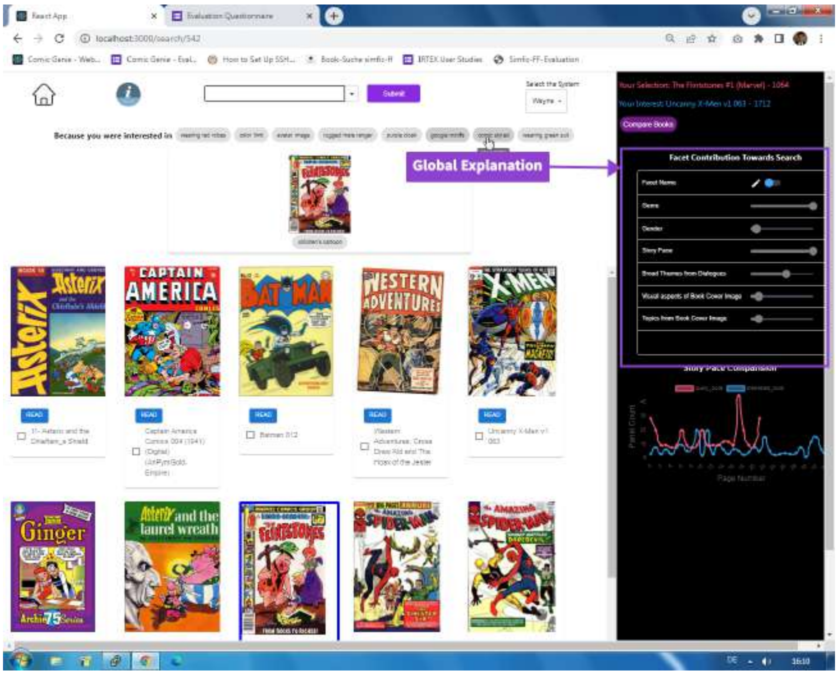
```
Require: $q$ = Query Book, Books hovered on by user : $D = h_1, h_2, ..., h_d$ treated as books of interest, Search Results
    from Low-Level Facet Ranker : $B = b_1, b_2, ..., b_m$ where $m = 200$ books
Ensure: Result after Re-Ranking $b'_1, b'_2, ..., b'_m \in B'$
Let $HistoricalSearchResults$ be list of search result books from past three turns.
Let $FACETWEIGHTS$ be updated facet weights from $UPDATEWEIGHTS$.
Let $VECTORIZE$ be a function that vectorize comic book into $k$ facet group.
$qv = VECTORIZE(q)$
for $i \leftarrow 1$ to $m$ books do
    $biv = VECTORIZE(bi)$
    Initialize facet_similarity score to 0, total_book_similarity to 0 and total_similarity_scores_list to [].
    for $j \leftarrow 1$ to $k$ facets do
        facet_similarity = cosine($qv, biv$) $* FACETWEIGHTS[j]$
        if $bi \in$ HistoricalSearchResults then
            facet_similarity = facet_similarity − 0.1
        end if
        total_book_similarity = total_book_similarity + facet_similarity
        Append total_book_similarity to total_similarity_scores_list
    end for
end for
Return $B'$ by sorting based on total_similarity_scores_list
```

Table 3.1: Facet Information

Facet Tier	**Facet Type**	**Facet Name**	**Dimensions**	**Rationale**
Low Level	Visual	CLD	64	Represent average colors in book, thereby style
		EHD	128	Edge representation in book, thereby style
		HOG	256	Object representations in book, thereby style
	Textual	Text - Bag of Words	64	Represent keywords from book
High Level	Visual	Book Cover Image	2048	Book cover image
		Book Cover Prompt	384	Prompts from book cover text
	Textual	Genre	16	Genre of the book
		Gender	3	Male oriented, Female oriented or others
		Supersense	46	Coarse categories of ontologies, helpful for question answering
	Temporal	Story Pace	1	Average panels per page representing rough pace
		Total Information in Book	1	Sum of text word count and panels in book

Step 4: Explanations for Search Results

Figure 3.10: Sliders weights as global explanation


React App
Evaluation Questionnaire
localhost:3000/search/542
Submit
Select the System
Wayne
Because you were interested in
Global Explanation
Your Selection: The Flintstones #1 (Marvel) - 1064
Your Interest: Uncanny X-Men v1 063 - 1712
Compare Books
Facet Contribution Towards Search
Facet Name
Genre
Gender
Story Pace
Broad Themes from Dialogues
Visual aspects of Book Cover Image
Topics from Book Cover Image
Story Pace Comparison
Panel Count
Page Number
READ

Global explanation pinpoints emphasized facets that retrieved top-K books, for a particular query book. Normalised facet weights from the optimisation algorithm 1 acts as emphasis as shown in fig 3.10. *Slider* components are used to implement weights in UI. *Slider* acts as two-way switch. More the slider weight value, higher the emphasis of facet towards search. User can enable system's default personalization or can input their own preferences through sliders 3.10 enabling interaction with the system.

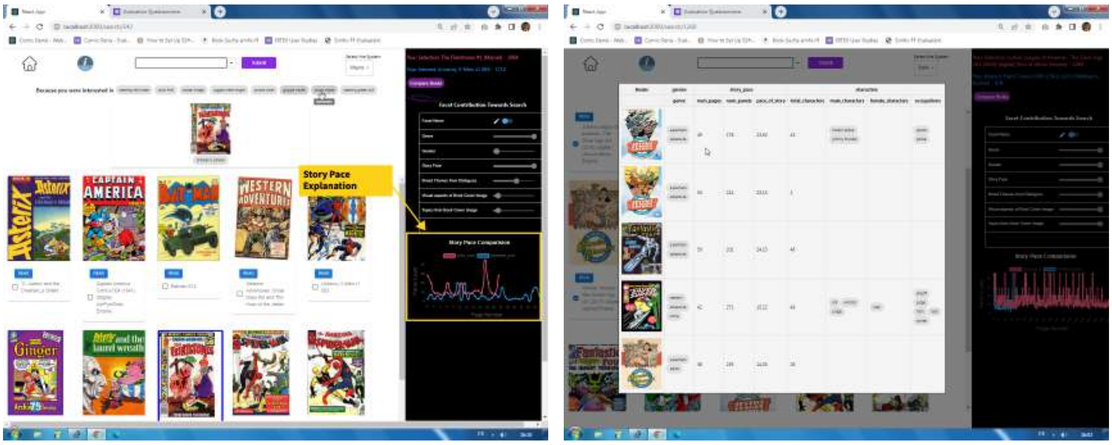
Local explanations describes the relationship between a given book and the query based on facets. A visual line-chart component depicts story pace between query and hovered book. On hover, We also display important topics related to persons, location, actions from the book as shown in 3.11. Topics are different than facets. We use BookNLP (Bamman et al. 2014) [43] to extract topics using NER on each book dialogue text. Users can also compare multiple books with table as shown in 3.11.

Users may wonder, *How did system came to conclusion on their preferences?*. We answer this through book cover prompts, as they can be easily verified against book cover

[43] https://github.com/booknlp/booknlp

image. Book prompt from current results are compared to hovered books in previous turn. Current book can be linked to previous book aspects. We encode book prompts by a sentence transformer [164] with model *all-MiniLM-L6-v2*. Encodings are compared against the corpus of previous turn hovered books using *cosine similarity*. The results are ranked and top 6 best pairs are displayed to the user. Explanation displayed on UI as shown in fig 3.12. Logic is represented by below algorithm 4.

Figure 3.11: Local explanations of AESC



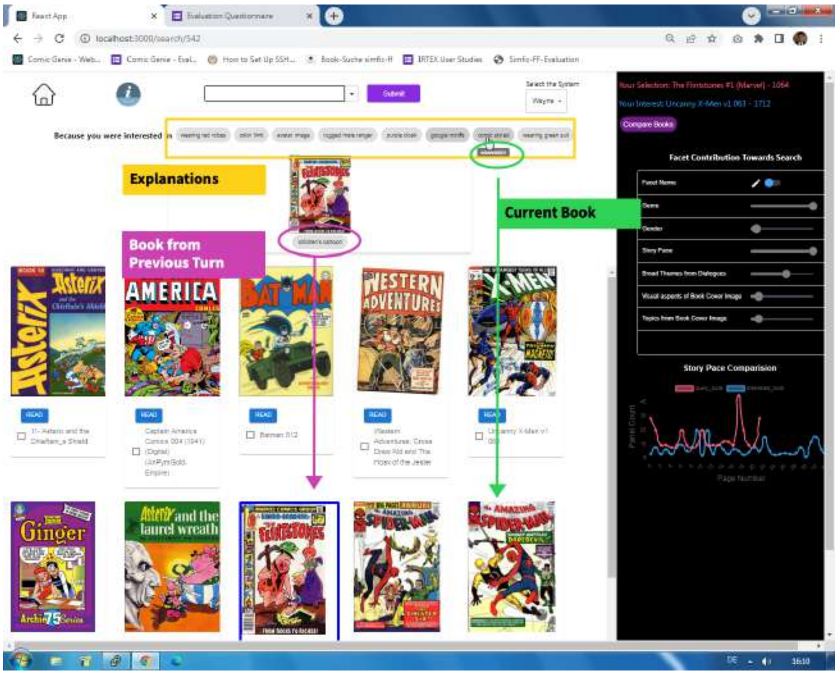
(a) Story pace comparison between query and hovered book in AESC

(b) Comparison of information between multiple selected books in AESC

Algorithm 4 Explain Personalization

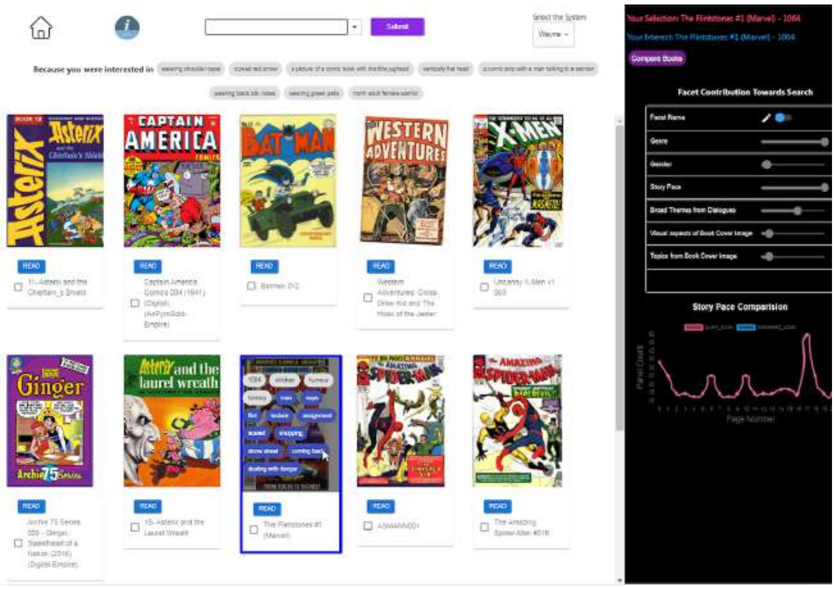
```
Require: LPH signify list of prompts of hovered book covers of previous turn
Require: CPH signify list of prompts of current book covers
LPH and CPH are lists of lists, where each sub-list has multiple prompts of a comic book cover
Ensure: Top-6 similarity scores and corresponding encoding pairs
Initialization: similarity_scores = [], encoding_pairs = []
Compare embeddings:
for bookPromptsListHistorical in LPH do
    for bookPromptsListCurrent in CPH do
        for bookPromptHistorical in bookPromptsListHistorical do
            for bookPromptCurrent in bookPromptsListCurrent do
                similarity = cosine_similarity(bookPromptHistorical, bookPromptCurrent)
                similarity_scores.append(similarity)
                encoding_pairs.append((bookPromptHistorical, bookPromptCurrent))
            end for
        end for
    end for
end for
Sort with similarity score and return top-6 previous book and current book prompt pair
Return top-6 previous book and current book prompt pair
```

Figure 3.12: Textual explanation and its nearest link to book from previous turn in AESC


React App
Evaluation Questionnaire
localhost:3000/search/542
Comic Genie - Web...
Comic Genie - Eval...
How to Set Up SSH...
Book-Suche simfic-ff
IRTEX User Studies
Simfic-FF-Evaluation
Submit
Select the System
Wayne
Because you were interested in
Explanations
Book from Previous Turn
Current Book
Your Selection: The Flintstones #1 (Marvel) - 1064
Your Interest: Uncanny X-Men v1 063 - 1712
Compare Books
Facet Contribution Towards Search
Facet Name
Genre
Gender
Story Pace
Broad Themes from Dialogues
Visual aspects of Book Cover Image
Topics from Book Cover Image
Story Pace Comparision
Panel Count
Page Number
READ
11- Asterix and the Chieftain_s Shield
Captain America Comics 004 (1941) (Digital) (AnPymGold-Empire)
Batman 012
Western Adventures: Cisco Drew Kid and The Hoax of the Jester
Uncanny X-Men v1 063

Step 5: Present Search Results

Figure 3.13: User Interface of AESC


Submit
Select the System
Wayne
Because you were interested in
Your Selection: The Flintstones #1 (Marvel) - 1064
Your Interest: The Flintstones #1 (Marvel) - 1064
Compare Books
Facet Contribution Towards Search
Facet Name
Genre
Gender
Story Pace
Broad Themes from Dialogues
Visual aspects of Book Cover Image
Topics from Book Cover Image
Story Pace Comparision
Panel Count
Page Number
READ
11- Asterix and the Chieftain_s Shield
Captain America Comics 004 (1941) (Digital) (AnPymGold-Empire)
Batman 012
Western Adventures: Cisco Drew Kid and The Hoax of the Jester
Uncanny X-Men v1 063
Archie 75 Series 005 - Ginger, Sweetheart of a Nation (2016) (Digital-Empire)
15- Asterix and the Laurel Wreath
The Flintstones #1 (Marvel)
The Amazing Spider-Man #016

We use a map based user interface 2.6.2 [39] as it provides minimal bias against ranking list. It allows observation of more books, as natural order is removed. However, to avoid user fatigue we limit books to 14. Fig 3.13 shows visual of map based UI.

4

Evaluation through User Study

This chapter begins with an overview of the user study, study design, followed by setup and a detailed presentation of the evaluation results. We closely follow user study experiment methodology from Kelly et al. [85] and Liu et al. [165]. Section ends with a summarization of the major takeaways from the evaluation. Our evaluation form can be accessed at link [44]

4.1 Methodology

4.1.1 Hypothesis

RQ1 : What is the impact of domain based facets on search results quality?

AESC is built on top of domain based facets, this naturally should improve the number of relevant documents in the search results. Increased relevance should make user more interested and satisfied. Interest leads to more interaction with the system. More the interaction could increase mental load too. Would using domain based facets improve the search quality of SERP?

RQ2 : What is the impact of user-system interaction on satisfying user's search context?

Every user could have their own goals while issuing a search query. Some user may be looking for shorter running time, some for similar characters other for genre. It becomes important to allow users to communicate their needs to the system. This is made possible through sliders that directly set facet weights of similarity model. Naturally, explicit supply of context should improve the number of relevant documents in the SERP. If users do not see good results, they might try to continuously test with

[44] `https://forms.gle/ZPGGHf3NywhnYCuW9`

different facet weights increasing interaction. Would interaction with sliders help users effectively communicate with the system to increase the search quality of SERP?

RQ3 : How does textual explanations generated from book cover help user understand personalization?

Book cover is the first visual of the comic and helpful to attract users to pick it up. AESC uses previous hovered books to personalize current results for a query. Linking elements of matching book covers between current search results and previous hovered books could uncover the personalization link. For example: a dress color, action scene, and objects in the scene. Naturally, we use textual prompts describing book cover as explanations. If explanations are of good quality, then they must explain a relevant connection between current and previous book. Users may interact more with texts that explain better. Are such explanations useful for users to understand the link between their previous hovered book and current search results, thereby personalization?

RQ4 : How does comparison table explanation help user discern between books from search results?

Users may need to discern between many books in SERP. It could help them to find the exact books of their need from bunch of result books. Comparison table is used [72] to differentiate between books. User can make use of comparison table to find books that fits their need. If they can find their books easily, then interaction must reduce for AESC. For example: finding books belonging to action genre. Would users be successful in finding books that fit their needs using comparison table?

4.1.2 Research Focus and Variables

Tasks are designed to answer each research question. We manipulate independent variable corresponding to each task and measure the dependent variable that quantifies change. Quasi-independent variables such as gender, domain expertise are collected. This is summarized in below table 4.1.

RQ1 : What is the impact of domain based facets on search results quality?

We aim to measure search quality of AESC system against baseline. To measure search quality, we should allow user to freely explore the system. Exploration is controlled

Table 4.1: Research Focus and Employed Variables

Research Question	**RQ1**	**RQ2**	**RQ3**	**RQ4**
Task Complexity	Controlled Exploratory	Controlled Exploratory	Fact-Finding	Fact-Finding
Research Focus	System Evaluation (Retrieval Performance)	System Evaluation (Interaction)	System Evaluation (Personalization Explanation)	System Evaluation (Comparison Table)
Independent Variables	Facets	Facet Weights	Textual Explanations	Story Pace in Comparison Table
Dependent Variable - Performance Measures	Interactive Precision	Interactive Precision	-	-
Dependent Variable - Usability Measures	Effectiveness, Efficiency and Satisfaction	Effectiveness, Efficiency and Satisfaction	Effectiveness, Efficiency and Satisfaction	Effectiveness, Efficiency and Satisfaction
Dependent Variable - Meta Measures	Cognitive Load, Hovered Books, Textual Explanations Clicked	Cognitive Load, Hovered Books, Textual Explanations Clicked	Cognitive Load, Hovered Books, Textual Explanations Clicked	Cognitive Load, Hovered Books, Textual Explanations Clicked
Quasi-Independent Variables	Gender, Domain Expertise, Computer Science Background	Gender, Domain Expertise, Computer Science Background	Gender, Domain Expertise, Computer Science Background	Gender, Domain Expertise, Computer Science Background

by making user choose pre-defined initial query and limiting turns. Furthermore, search quality should increase across subsequent turns. We judge performance of AESC against baseline using interactive precision (IP), and usability metrics. We observe user activity through meta measures to get a bigger picture.

RQ2 : What is the impact of user-system interaction on satisfying user's search context?

We aim to measure change in search quality of AESC system against baseline by interaction. To measure search quality, we give user a typical day-day requirement in plain english and let user make the decision of changing appropriate sliders. For example: You are interested in action packed short movies. Furthermore, search quality should increase after user's input. Exploration is controlled by making user choose pre-defined initial query and limiting turns. We judge performance of AESC against baseline using IP, correct sliders used, and usability metrics. We observe user activity through meta measures to get a bigger picture.

RQ3 : How does textual explanations generated from book cover help user understand personalization?

We would like to check, if textual explanations derived from previous activity and current SERP are useful to explain personalization for the user. To measure explanation usefulness, we design a fact-finding task, make use select a predefined query and choose the useful explanation of their choice. We judge performance of AESC against baseline through usability metrics. We observe user activity through meta measures to get a bigger picture.

RQ4 : How does comparison table explanation help user discern between books from search results?

We would like to check, if comparison table is useful for the user to differentiate between books from SERP. We are also not looking to compare comparison table against other explanations. To measure explanation usefulness, we design a fact-finding task, make user select books that have specific characterstics from a pre-defined SERP. We judge performance of AESC against baseline through usability metrics. We observe user activity through meta measures to get a bigger picture.

4.1.3 Metrics

Contextual Metrics

We collect users age, name and their background in computer science. User's experience in computer science and information retrieval can make it a tad bit easy to do the task. We also collect their comic background, whether they are readers, artists or non-readers. Fans usually know books at back of their hand, Hence can take less time in assessing relevance of the result book.

Performance Metrics

IP is the proportion of relevant documents recorded by the user to the total number of documents recorded in SERP. It captures the proportion of selected documents that are actually relevant to the user's information needs 4.1.

$$\text{Interactive Precision} = \left(\frac{\text{Number of Relevant Documents Selected}}{\text{Total Number of Documents Selected}}\right) \times 100 \tag{4.1}$$

Interactive recall is the proportion of the number of relevant documents recorded by the user to the total number of relevant documents in the corpus. It measures the effectiveness of the system in retrieving all relevant documents based on the user's interactions 4.2.

$$\text{Interactive Recall} = \left(\frac{\text{Number of Relevant Documents Selected}}{\text{Total Number of Relevant Documents in the Corpus}}\right) \times 100 \tag{4.2}$$

Both metrics take into account the user's actions and judgments during the search process, providing a more dynamic and user-centric evaluation of the system's performance compared to traditional precision and recall measures.

Interaction Metrics

We extract interaction metrics through user logs. We extract hovered books, time spent nd textual explanations checked. Clicks on SERP is not used as clicking means choosing query. Number of clicks are controlled by task. More hovered books indicate more time spent, explanations checked for the task could indicate positive interactions. More time spent could also mean more cognitive load which has to be validated through eye-tracker. Time is tracked in efficiency metric under usability umbrella.

Usability Metrics

Under usability umbrella, We measure effectiveness, efficiency and satisfaction for every task. Effectiveness of system in performing task indicates good fit for the task [166], as defined by equation 4.3. Taking less time in solving a task could mean that, the task could be performed faster in the system. Time can be extracted from logs.

$$\text{Task Effectiveness} = \frac{1}{100} \times (\text{Quantity} \times \text{Quality})\% \tag{4.3}$$

In this formula, Quantity represents the total number of correct artifacts present, and Quality represents the count of correct artifacts marked by the user. The task effectiveness is calculated by multiplying Quantity and Quality, dividing it by 100, and expressing it as a percentage. We tweak the formula to fit our purpose as mentioned below 4.4. Artifact can be a book or an explanation depending on the task.

$$\text{Task Effectiveness} = \left(\frac{\text{Number of Relevant Artifacts Chosen by User}}{\text{Total Relevant Artifacts}}\right) \times 100 \tag{4.4}$$

Satisfaction is something subjective to every user, Subjective opinions gives overall picture of the system. It is elicited from user and their response is recorded on likert scale of 1-7. Metrics are reported by averaging users feedback to mean μ and standard deviation σ. We use two-tailed Wilcoxon Signed Rank test to assess AESC against baseline [167] [140].

4.2 Design

Due to non-availability of comic book retrieval dataset benchmarks, We choose to evaluate retrieval and explanations through user study. We conduct a controlled within-subject user study, to answer our research questions under umbrella of usability metrics. We use Graeco-Latin square design to randomize task order as well as system order for different users. User needs to repeat the same task across both AESC and baseline as shown in table 4.2.

4.2.1 Setup

Tobii Eyetracker T60 was used as an eye-tracker. The user was given free movement of their head within a 44 x 22 x 30 cm range, which was mounted approximately 60

Table 4.2: Graeco-Latin Square Experiment Design

User	Task Order	System Order	Task	System
1	1	1	RQ1	AESC, Baseline
1	2	2	RQ2	Baseline, AESC
1	3	1	RQ3	AESC, Baseline
1	4	2	RQ4	Baseline, AESC
2	4	1	RQ4	AESC, Baseline
2	1	2	RQ1	Baseline, AESC
2	2	1	RQ2	AESC, Baseline
2	3	2	RQ3	Baseline, AESC
3	3	1	RQ3	AESC, Baseline
3	4	2	RQ4	Baseline, AESC
3	1	1	RQ1	AESC, Baseline
3	2	2	RQ2	Baseline, AESC
4	2	1	RQ2	AESC, Baseline
4	3	2	RQ3	Baseline, AESC
4	4	1	RQ4	AESC, Baseline
4	1	2	RQ1	Baseline, AESC

cm from them. It has a 17-inch TFT display with a 1280 x 1024 pixel resolution, with data rate of 60 Hz and an accuracy of 0.5 degrees. The eye tracker had an inaccuracy of roughly 0.5 cm at a distance of 60 cm between the user and the screen. To provide comfort and the right distance from the monitor, participants received an adjustable chair [168] [169]. Both eye-tracker computers and normal computers were used, users were assigned randomly to them. Nine users performed evaluation in eye-tracker The basic user study process is as follows.

1. User gets assigned eye-tracker or a normal computer for evaluation.
2. Users watch a video describing system and its features. We provide more details verbally to the user.
3. Users do a demographics survey comprising of contextual measures.
4. Users undergo free system exploration for 5-10 minutes to familiarize themselves.
5. Read the task and watch it's associated short video clip.
6. Perform the task, record the feedback and experience for the task.
7. Collect pain-points through free text to conclude search results

4.3 Evaluation Details

4.3.1 Participant Details

Users were invited through email from university emailing system. User study was controlled and conducted in a lab. Totally 33 users participated in user study and one user was omitted as user did not evaluate complete tasks, making it 32 for evaluation. There were 23 males and 10 females as shown in fig 4.1. 1 user frequently read comics, 7 hobby readers, 20 users have watched comic based movies but have not read books, 4 users have not read or watched comic based entertainment as shown in fig 4.1. We can deduce that most of our users are non-domain experts and only have general grounding in comics. This may impact evaluation, if users do not see some famous comics that they have seen in day-day movies. On computer science background, 27 users used search engines like google extensively, 16 users have studied information retrieval, 28 students and 4 from research background 4.1. We can deduce that most of our users are computer savvy and have good grounding in google search. All four researchers had information retrieval background. Study had 17 computer-science users and 15 non computer-science users 4.1. Hence it was almost evenly distributed. Almost half of the users are not from computer science background, this could make interactive search a bit complicated. 10 users evaluated on eye-tracker system, whereas 23 users evaluated on normal computers.

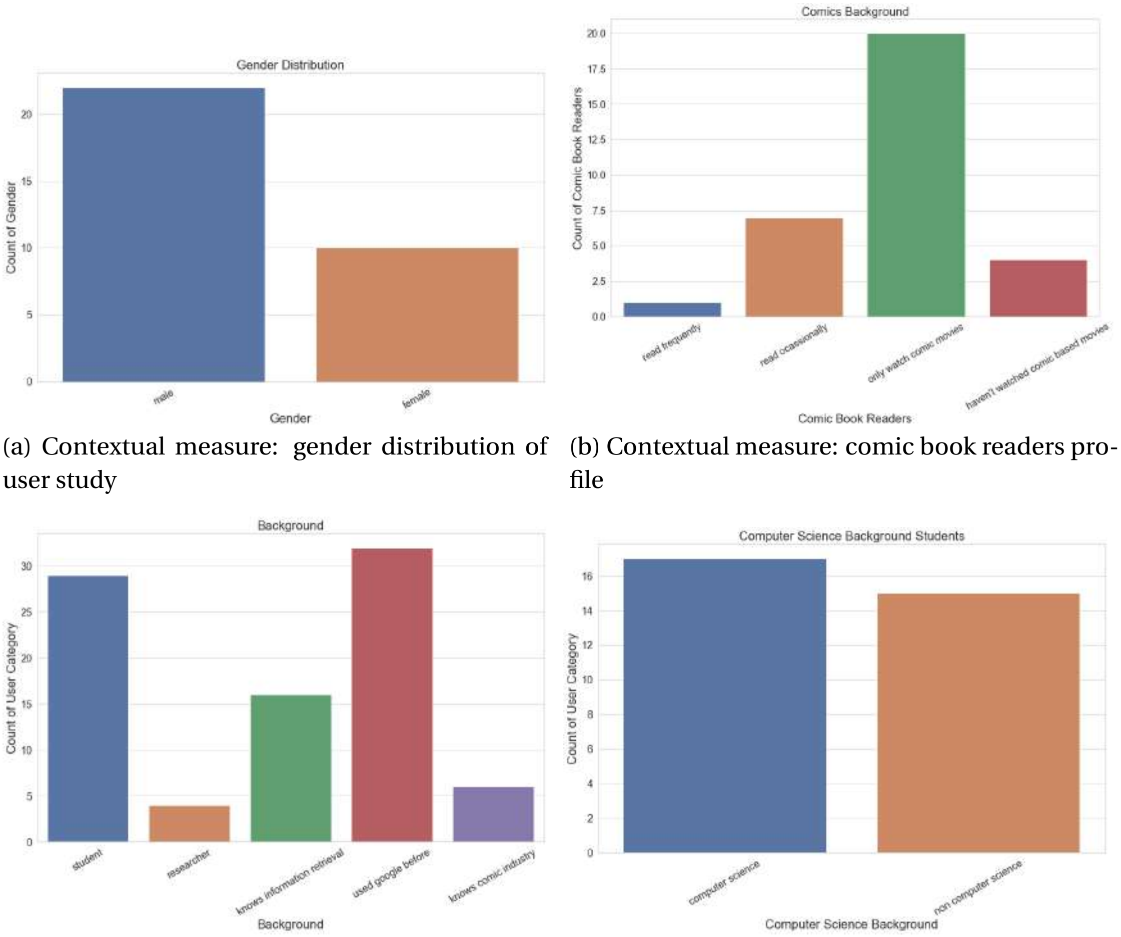
4.3.2 RQ 1: What is the impact of domain based facets on search results quality?

We attempt to answer, RQ1through a task against both AESC and baseline. We ask user to perform task on both AESC and baseline. We record the interaction details and user's opinion.

Baseline System

We create a baseline system that has identical user interface as AESC. To evaluate choice of facets, We need to compare against another system with different facets or no facets. Since, We do not have benchmark system with different facets, we attempt to compare AESC to a pseudo-random retriever. Facet weights are also changed randomly to provide illusion of adaptation.

Figure 4.1: Contextual details about participants



(a) Contextual measure: gender distribution of user study

(b) Contextual measure: comic book readers profile

(c) Contextual measure: comic book readers background

(d) Contextual measure: proportion of computer science comic book readers

Task

We ask user to select book “Cowgirl Romances” or “Jumbo Comics - Sheena Queen of the Jungle” from landing page. We already know the relevant result list for these books through manual inspection. The books are selected in such a way that, there are atleast 5 relevant books in turn 1 and turn 2. User records books that they feel relevant for the query. Later, they are free to choose any of the book as query from current result set. In next turn, user again records relevant books in current result. We cross check users relevant books with our relevant books list and calculate effectiveness 4.4 and IP 4.1.

We extract task time taken from logs. We also ask users to record their thoughts behind their choice, However this is not mandatory. Task is repeated across both systems.

User assigns a score between 1-7 on likert scale for satisfaction about the task performance in both systems. Baseline system could have lesser relevant books than AESC, as it is not retrieved through facets.

Discussion

Table 4.3 provides captured metrics for count of hovered books and explanations checked used at each turn for both systems. Due to open-ended nature of the task, We expected interaction with AESC would be significantly greater than baseline, alas the interactions with both systems were similar. Users hovered marginally more books from AESC ($\mu = 5.09$, $\sigma = 1.85$) against baseline ($\mu = 5.02$, $\sigma = 1.21$) but with no significance $p = 0.22$. This trend continues for textual explanations, users clicked on marginally lesser explanations from AESC ($\mu = 1.12$, $\sigma = 0.32$) versus baseline ($\mu = 1.23$, $\sigma = 0.31$) with no significance $p = 0.26$. From observing video logs, users tend to click on explanations only if they feel interested in the text or can connect it to the book cover. This is also in sync with explanation quality.

Table 4.3: RQ1: Comparison of Interaction Details

Turn	**AESC**		**Baseline**	
	Hovered Books	**Textual Explanation Clicked**	**Hovered Books**	**Textual Explanation Clicked**
1	**5.42 ± 1.98**	-	5.12 ± 1.44	-
2	4.76 ± 1.73	1.15 ± 0.32	**4.93 ± 0.98**	**1.23 ± 0.41**
overall	**5.09 ± 1.85**	1.15 ± 0.32	5.02 ± 1.21	**1.23 ± 0.41**

Table 4.4 provides captured usability metrics such as IP, effectiveness, efficiency and satisfaction at each turn for both systems. Comparing precision, we see that overall performance of AESC ($\mu = 77.23$, $\sigma = 37.91$) is better than baseline ($\mu = 52.07$, $\sigma = 49.66$) with statistical significance $p = .04136$. Furthermore, precision between turns for AESC increases from ($\mu = 70.2$, $\sigma = 41.71$) to ($\mu = 84.26$, $\sigma = 34.11$) with significance $p = 0.0065$. Although, performance of baseline from ($\mu = 51.1$, $\sigma = 50.1$) to ($\mu = 53.2$, $\sigma = 49.26$), it did not have significance $p = 0.71$. For the first turn, We observe that AESC was more effective ($\mu = 39$, $\sigma = 29.9$) against baseline ($\mu = 20.02$, $\sigma = 19.8$) with significance $p = 0.00039$. As user choose next book and entered turn 2, we observe that AESC effectiveness went significantly higher to ($\mu = 55.62$, $\sigma = 36.53$) with significance of $p = 0.039$. Baseline systems effectiveness too went up to ($\mu = 30.62$, $\sigma = 30.04$), but

did not have significance of $p = 0.054$. This is evident by the fact that the standard deviation for baseline is varying largely and users are choosing from random system. Overall effectiveness of AESC ($\mu = 47.5, \sigma = 26.39$) was much better than baseline ($\mu = 25.31, \sigma = 20.94$) and attained statistical significance $p = 0.00016$. Satisfaction was also much better ($\mu = 5.03, \sigma = 1.51$) versus baseline ($\mu = 3.56, \sigma = 2.05$) with $p = 0.00038$. Overall time required to complete the task for AESC ($\mu = 2.47$, $\sigma = 1.5$) was more against baseline ($\mu = 1.58$, $\sigma = 0.34$) with statistical significance of $p = 0.031$. We can correlate this with the above interaction table 4.3 and increased relevant books recorded for AESC. More the interaction and recorded books, more the time taken, which is directly inline with time.

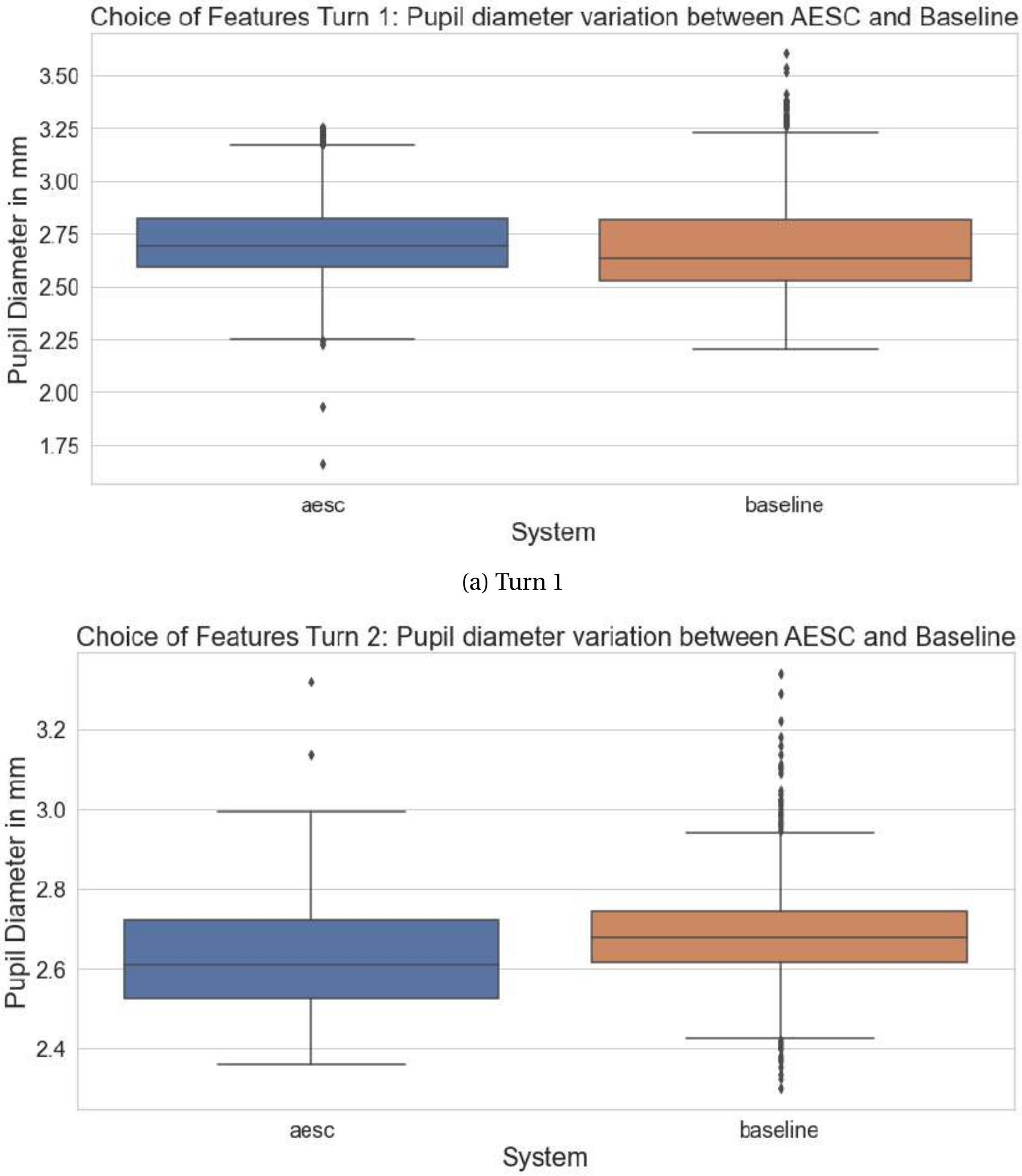
Table 4.4: RQ1: Comparison of Usability Metrics

Metric	**AESC**			**Baseline**		
	Turn 1	**Turn 2**	**Overall**	**Turn 1**	**Turn 2**	**Overall**
Interactive Precision (in %)	**70.2 ± 41.71**	**84.26 ± 34.11**	**77.23 ± 37.91**	51.1 ± 50.1	53.2 ± 49.26	52.07 ± 49.66
Effectiveness (in %)	**39 ± 29.39**	**55.62 ± 36.53**	**47.5 ± 26.39**	20.02 ± 19.8	30.62 ± 30.04	25.31 ± 20.94
Efficiency (in min)	2.61 ± 0.63	2.33 ± 1.5	2.47 ± 1.5	**1.7 ± 0.46**	**1.46 ± 0.23**	**1.58 ± 0.34**
Satisfaction	-	-	**5.03 ± 1.51**	-	-	3.56 ± 2.05

We compare cognitive load for AESC versus baseline and AESC across both turns as shown in fig 4.2. For first turn, it was evident that cognitive load was more for AESC ($\mu = 2.73, \sigma = 0.2$) than baseline ($\mu = 2.69, \sigma = 0.21$) with significance of $p = 0.0007$. However, AESC has lesser cognitive load for AESC ($\mu = 2.62, \sigma = 0.12$) than baseline ($\mu = 2.68, \sigma = 0.1$) with significance of $p = 0.0002$. This could be correlated to increased interactivity 4.3. As we enter turn 2, interactivity of AESC goes down as users see more relevant results and only interact them 4.4, whereas baseline has random results, hence user has to interact with more books and make tough decisions to select relevant books. This may have decreased cognitive load in turn 2. This is different than our hypothesis, where we thought that load should increase with interaction.

Examination of gaze plots and heatmaps for AESC versus baseline system show a similar picture for transition from first turn 4.3 to second turn 4.4. We see more interaction in turn 2 for baseline as users hover more to find details, whereas user interaction marginally decreases and remains similar to that of turn 1 for AESC. Finally, we can conclude that domain based facets has good retrieval performance for comics, can be enhanced in the future.

Figure 4.2: RQ1: Cognitive load variation between AESC and baseline



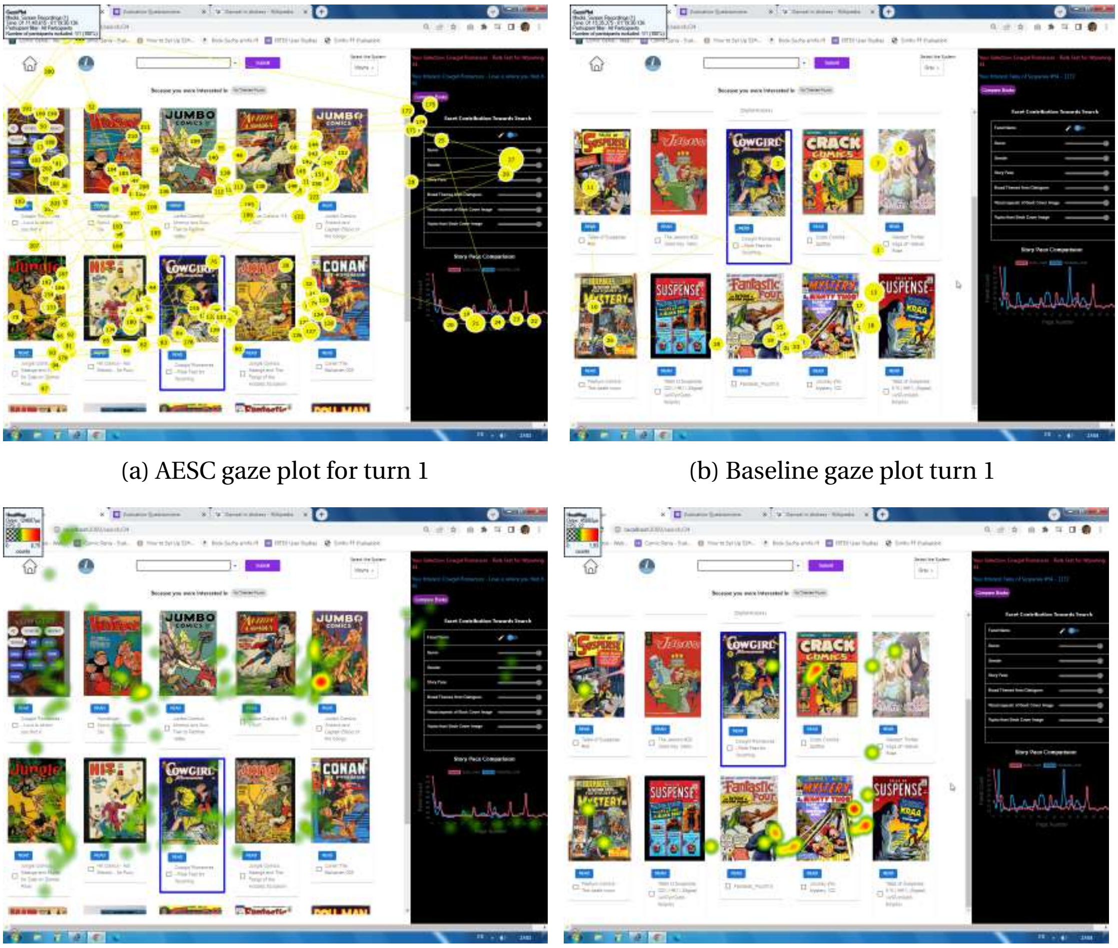
Choice of Features Turn 1: Pupil diameter variation between AESC and Baseline

(a) Turn 1

Choice of Features Turn 2: Pupil diameter variation between AESC and Baseline

(b) Turn 2

Figure 4.3: RQ1: Gaze plot and heat map variations between AESC and Baseline for Turn 1



(a) AESC gaze plot for turn 1

(b) Baseline gaze plot turn 1

(c) AESC heat map turn 1

(d) Baseline heat map turn 1

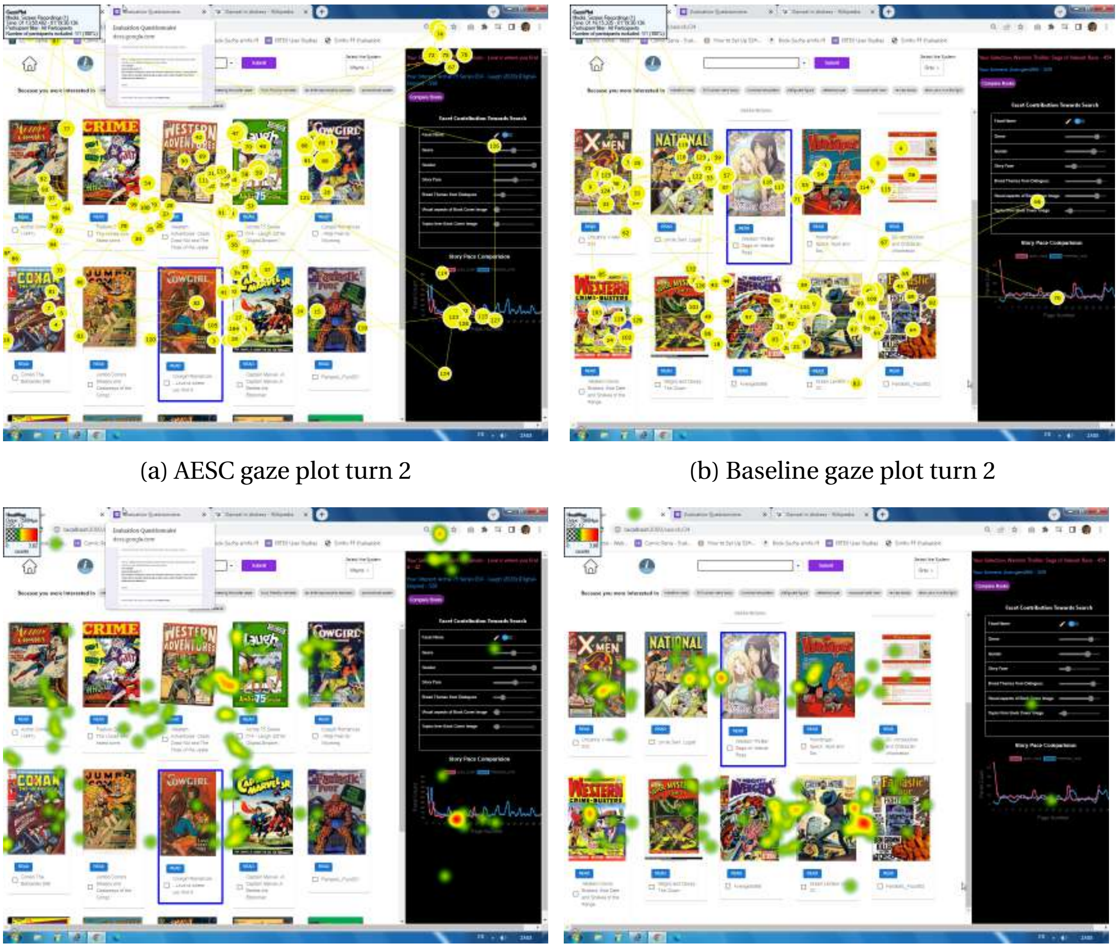
4.3.3 RQ 2: What is the impact of user-system interaction on satisfying user needs?

We attempt to answer, RQ2 through a task against both AESC and baseline. We ask user to perform task on both AESC and baseline. We record the interaction details and user's opinion.

Baseline System

User-System interaction is the core underpinning of AESC. For interaction to work two things must happen, System should be capable of personalizing to user context or user should be able to communicate their preference directly to the system. Failing so, will

Figure 4.4: RQ1: Gaze plot and heat map variations between AESC and baseline for turn 2



(a) AESC gaze plot turn 2

(b) Baseline gaze plot turn 2

(c) AESC heat map turn 2

(d) Baseline heat map turn 2

deem system incapable of interaction. We can still evaluate a system interaction and another system with no interaction, but this may bias user to think that interactive system is better or worse than the other. To evaluate such interaction we, neutralize adaptation algorithm with constant weights, remove re-ranker feedback from UI. We create a baseline system that has identical UI as AESC to remove any bias. Facet weights displayed to user also changes randomly to provide illusion of adaptation. Baseline system slider displays facet weights with random values. Sliders are still changeable showing an illusion of inter-activeness with the system, but their feedback is disconnected to the adaptation algorithm.

Task

User interacts with the system for two turns. They are indirectly elicited to change the relevant slider to reach the tasks goal. We provide a task goal of choosing sliders that could reflect "damsel-in-distres", "fast-paced" books without worrying much about "dialogues". Here, we ask user to select book "Jumbo Comics - Sheena Queen of the Jungle" from landing page at first turn. User records books that they feel relevant for the query. They provide facet weights by changing appropriate sliders that they think, best suited for achieving task goal. In next turn, they click on same query book to re-retrieve books. User again records relevant books in current result. We cross check users relevant books with our relevant books list and calculate effectiveness 4.4 and IP 4.1. We extract task time taken from logs. We also ask users to record their thoughts behind their choice, However this is non-mandatory. Task is repeated across both systems.

User assigns a score between 1-7 on likert scale for satisfaction about the task performance in both systems. Baseline system could have lesser relevant books than AESC, as it is not retrieved through facets.

Discussion

Table 4.5 provides captured metrics for count of hovered books and explanations checked used at each turn for both systems. Due to slight open-ended nature of the task, We expected interaction with AESC would be significantly greater than baseline, although the interaction was greater with out system, it was not uniform. Users hovered more books from AESC ($\mu = 5.52$, $\sigma = 1.47$) against baseline ($\mu = 4.63$, $\sigma = 0.88$) with significance $p = 0.042$. The interaction also increased between turns for both the systems. On the contrary for textual explanations, users clicked on marginally lesser explanations from AESC ($\mu = 0.34$, $\sigma = 0.31$) versus baseline ($\mu = 0.35$, $\sigma = 0.14$) with no significance $p = 0.37$. From observing video logs, users tend to click on explanations only if they feel interested in the text. They also tend to interact by seeing book cover.

Table 4.5: RQ2: Comparison of interaction details

Turn	**AESC**		**Baseline**	
	Hovered Books	**Textual Explanation Clicked**	**Hovered Books**	**Textual Explanation Clicked**
1	**5.33 ± 1.57**	-	4.35 ± 0.78	-
2	**5.69 ± 1.37**	0.34 ± 0.31	4.93 ± 0.98	**0.35 ± 0.14**
overall	**5.52 ± 1.47**	0.34 ± 0.31	4.63 ± 0.88	**0.35 ± 0.14**

Table 4.6 provides captured usability metrics such as IP, effectiveness, efficiency and satisfaction at each turn for both systems. Comparing precision, we see that overall performance of AESC ($\mu = 75.71, \sigma = 31.44$) is better than baseline ($\mu = 61.75, \sigma = 43.87$) with statistical significance $p = .034$. Furthermore, precision between turns for AESC increases from ($\mu = 63.72, \sigma = 35.95$) to ($\mu = 87.7, \sigma = 26.92$) with significance $p = 0.048$. Although, performance of baseline from ($\mu = 59.9, \sigma = 45.1$) to ($\mu = 63.62, \sigma = 42.26$), it did not have significance $p = 0.73$. For the first turn, We observe that AESC was similar ($\mu = 682.1, \sigma = 28.6$) to baseline ($\mu = 68.1, \sigma = 30$) with no significance $p = 0.89$. This was expected as both systems produce same output for first turn. As user choose next book and entered turn 2, we observe that AESC effectiveness went marginally lower higher to ($\mu = 67.5, \sigma = 33.8$) with no significance $p = 0.95$. This was not expected as we thought that, effectiveness would go up in turn 2 correlating with interactions and adaptability. Baseline systems effectiveness too went significantly down to ($\mu = 53.12, \sigma = 31.15$), but did not have significance of $p = 0.054$. This was expected as baseline system has no adaptability. Overall effectiveness of AESC ($\mu = 67.81, \sigma = 26$) was much better than baseline ($\mu = 60.62, \sigma = 23.4$), but had no statistical significance $p = 0.062$. Satisfaction was much better ($\mu = 5.41, \sigma = 1.36$) versus baseline ($\mu = 4.43, \sigma = 1.72$) with $p = 0.023$. Overall time required to complete the task for AESC ($\mu = 7.01, \sigma = 5.63$) was more against baseline ($\mu = 3.16, \sigma = 2.45$) with statistical significance of $p = 0.0048$. From logs, it was observed that, users spent more time playing around with sliders and checking the results against baseline. This resulted in more time for AESC. This tally well with assumed user fatigue.

Table 4.6: RQ2: Comparison of usability metrics

Metric	AESC			Baseline		
	Turn 1	Turn 2	Overall	Turn 1	Turn 2	Overall
Interactive Precision (in %)	**63.72 ± 35.95**	**87.7 ± 26.92**	**75.71 ± 31.44**	59.9 ± 45.1	63.62 ± 42.26	61.75 ± 43.87
Effectiveness (in %)	**68.12 ± 28.67**	**67.5 ± 32.82**	**67.81 ± 26**	68.12 ± 30.04	53.12 ± 31.15	60.62 ± 23.41
Efficiency (in min)	6.4 ± 4.61	7.62 ± 6.65	7.01 ± 5.63	**2.5 ± 0.61**	**3.83 ± 1.81**	**3.16 ± 2.45**
Satisfaction	-	-	**5.41 ± 1.36**	-	-	4.43 ± 1.72

We compare cognitive load for AESC versus baseline and AESC across both turns 4.5. For first turn, it was surprising that cognitive load varied so much provided same results were displayed for first turn across both the systems. AESC has lesser cognitive load ($\mu = 2.45, \sigma = 0.11$) than baseline ($\mu = 2.74, \sigma = 0.17$) with significance $p = 0.0001$.

However, AESC has more variations at higher end of pupil diameter indicating that, users may have been surprised to see same results in both systems for first turn. For second turn, It is clear that cognitive load of AESC ($\mu = 3.1$, $\sigma = 0.08$) shoots up above baseline ($\mu = 2.9$, $\sigma = 0.04$), but it has no significance $p = 0.25$. This could be correlated to increased interactivity 4.5. As AESC is adaptable, users see more relevant books 4.6 and become more interactive increasing load. This is opposite to our hypothesis, that load should increase for baselines due to bad results.

Examination of gaze plots and heatmaps for AESC versus baseline system show a similar picture for transition from first turn 4.6 to second turn 4.7. We see more interaction in slider component and decreased focus on search results. This is goes well with interaction metrics 4.5. Finally, we can conclude that users can interact with the system and convey their thoughts, albeit with lot of fatigue.

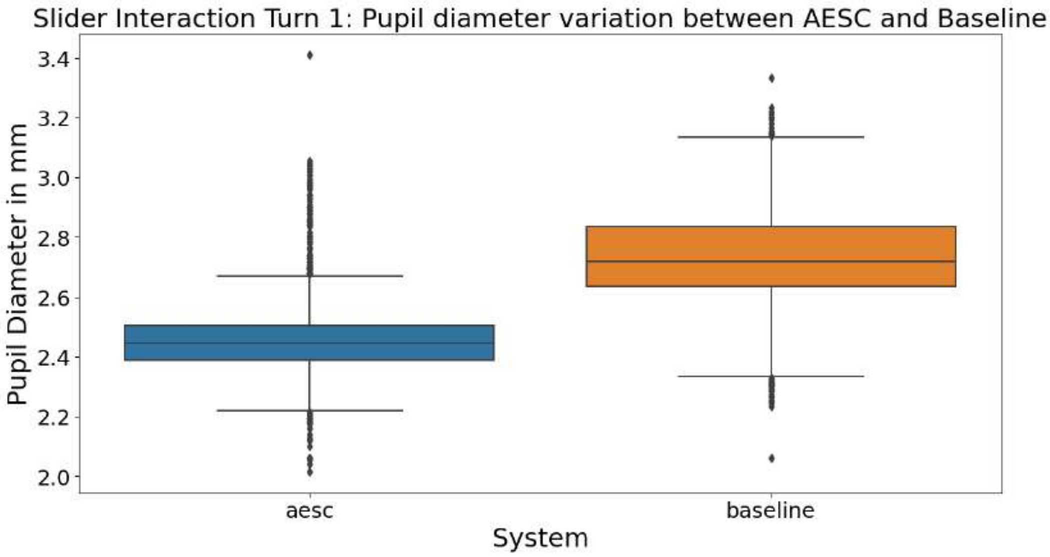
4.3.4 RQ 3: How does textual explanations generated from book cover help user understand personalization?

We attempt to answer, RQ3 through a task against both AESC and baseline. We ask user to perform task on both AESC and baseline. We record the interaction details and user's opinion.

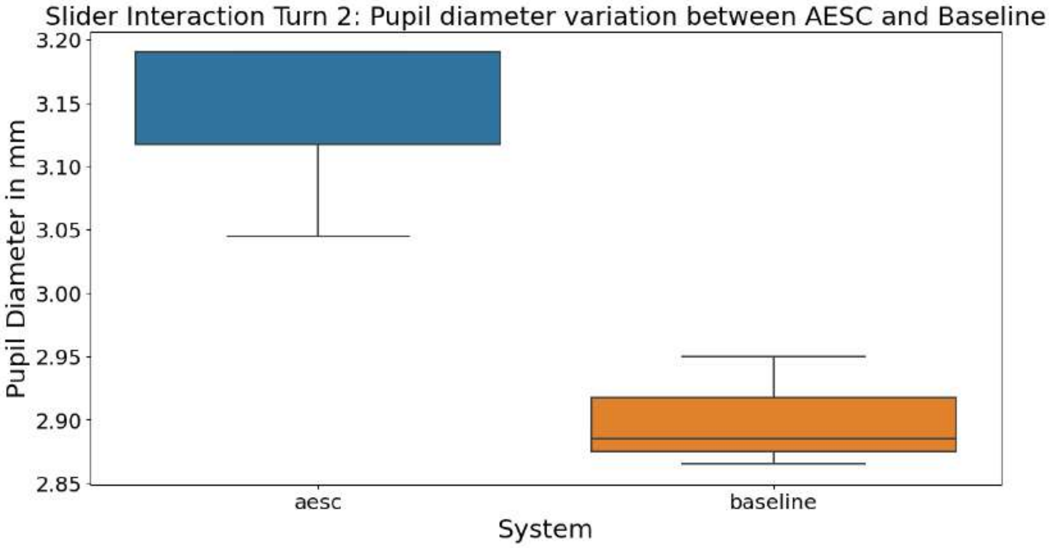
Baseline System

Explaining personalization from user-system interaction is one of the core underpinning of AESC. AESC uses interaction from previous turn to adapt results from current turn. Hence for explanations to be useful, It should be able to make user understand connection between their current search results and their interaction in previous turn. Failing to do so, will deem system incapable of explaining interactions. We can still evaluate a system with interaction explanation against another system with no interaction explanation, but this may bias user to think that system with explanation is better than its counterpart. To evaluate such explanation, we randomize textual explanations. We create a baseline system that has identical UI as AESC to remove any bias. Random textual explanations is provided and they are associated with their previous turn book to make it indistinguishable against AESC. We also retain personalization to provide same results as AESC, so as to not bias user using search results.

Figure 4.5: RQ2: Cognitive load variation between AESC and baseline



(a) Turn 1

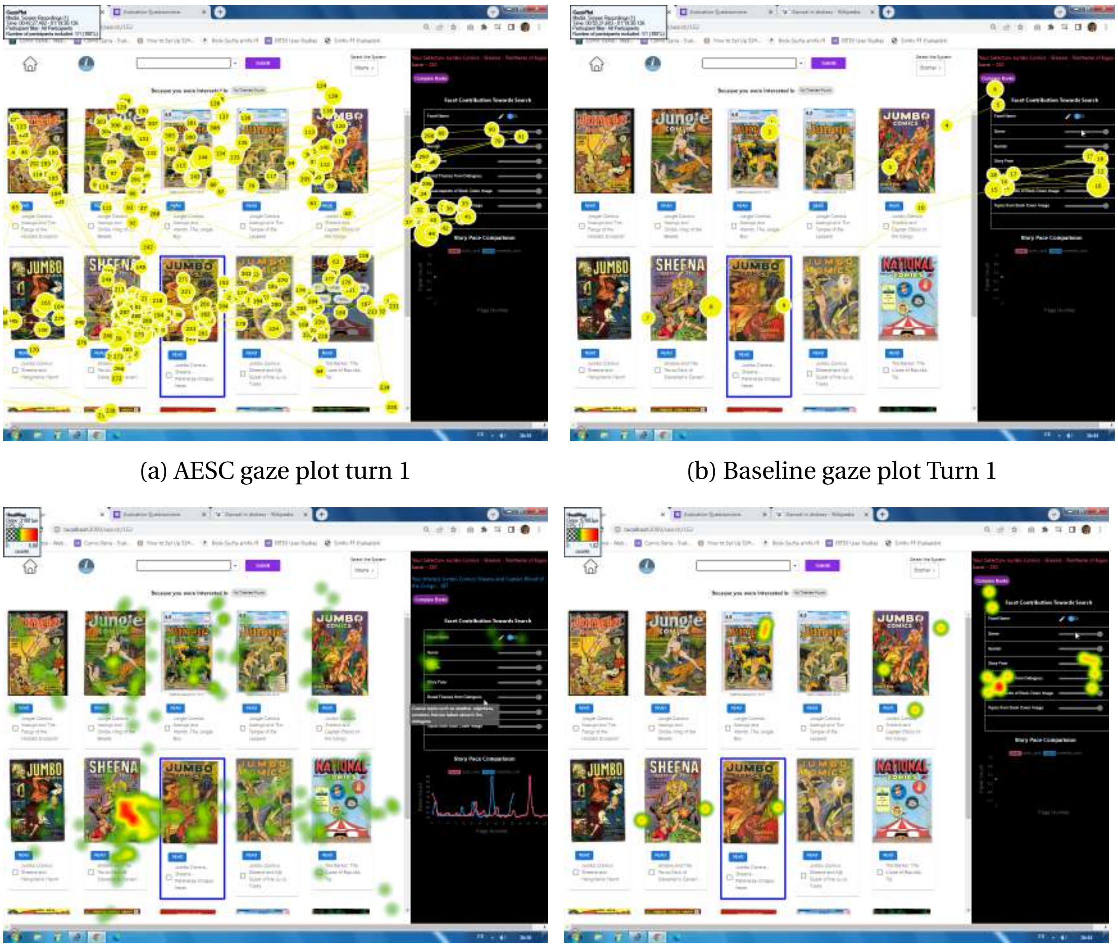


(b) Turn 2

Task

User interacts with the system for two turns. They are prodded to interact appropriate books to achieve task goals. In first turn, user chooses 'Asterix' book from landing page. Then they interact with first five books on screen, observing local explanations. In

Figure 4.6: RQ2: Gaze plot and heat map variations between AESC and baseline for turn 1



(a) AESC gaze plot turn 1

(b) Baseline gaze plot Turn 1

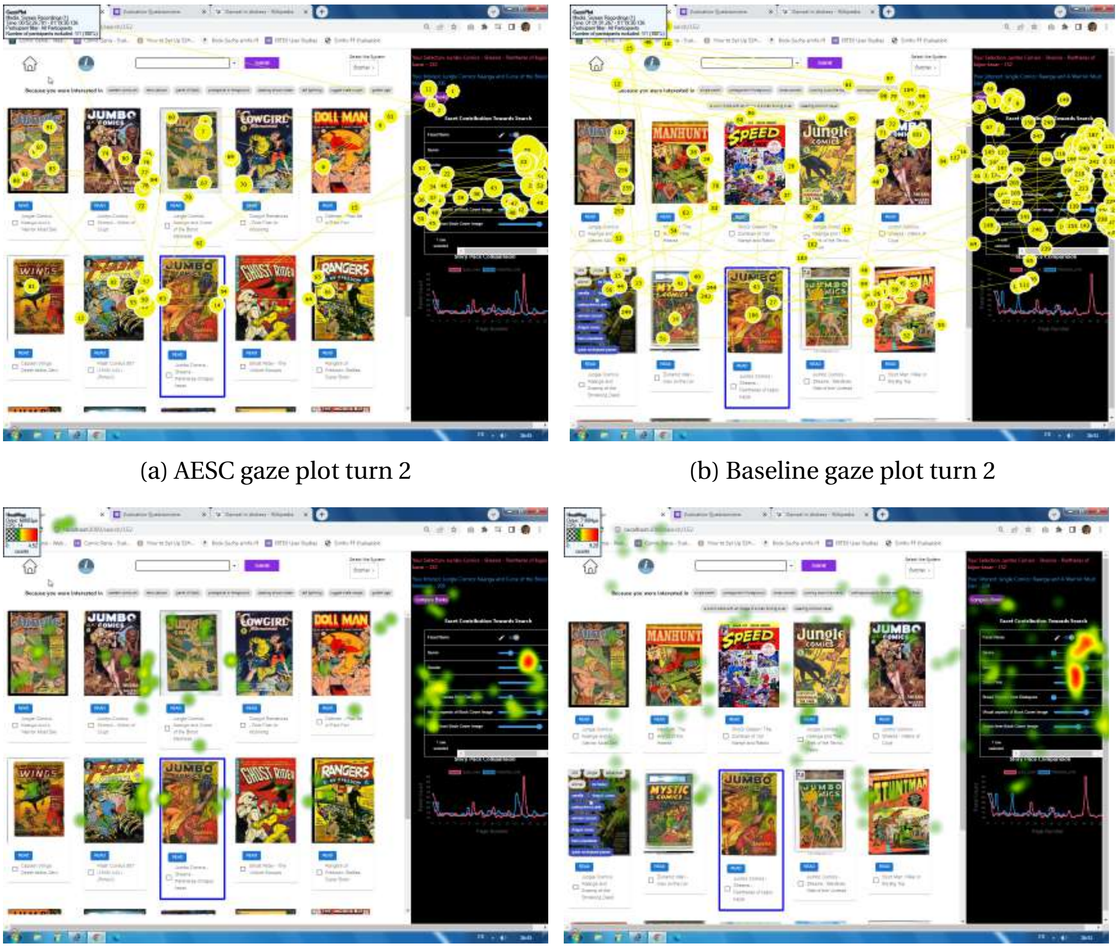
(c) AESC heat map Turn 1

(d) Baseline heat map turn 1

second turn, they click on 'Flintsones #1' and observe presented explanations on the page. If they like an explanation, then they can click on it to check how it is directly connected to a book from previous turn. Users record useful explanations that explain both current book and its link to book from previous turn. We cross check users explanations with our relevant explanations list and calculate effectiveness 4.4. We also ask users to record their thoughts behind their choice, However this is non-mandatory. Task is repeated across both systems.

User assigns a score between 1-7 on likert scale for satisfaction about the task performance in both systems. Baseline system could have lesser relevant books than AESC, as it is not retrieved through facets.

Figure 4.7: RQ2: Gaze plot and heat map variations between AESC and baseline for turn 2



(a) AESC gaze plot turn 2

(b) Baseline gaze plot turn 2

(c) AESC heat map turn 2

(d) Baseline heat map turn 2

Discussion

Table 4.7 provides captured metrics for count of hovered books, time taken, explanations checked and search-bar used at each turn for both systems. We deduce that, interaction has increased due to multiple turns and requirement of the task to evaluate explanations. However, users hovered more books from AESC ($\mu = 7.83$, $\sigma = 2.78$) against baseline ($\mu = 6.58$, $\sigma = 1.27$) but with no significance $p = 0.19$. This trend continues for explanations, users clicked on more explanations from AESC ($\mu = 5.52$, $\sigma = 1.03$) versus baseline ($\mu = 4.16$, $\sigma = 1.31$) with significance $p = 0.046$. From observing logs, users tend to click on explanations only if they feel interested in the prompt or can connect to the book cover.

Table 4.7: RQ3: Comparison of interaction details

AESC		Baseline	
Hovered Books	**Explanations Checked**	**Hovered Books**	**Explanations Checked**
7.83 ± 2.78	**5.52 ± 1.03**	6.58 ± 1.27	4.16 ± 1.31

Table 4.8 provides captured usability metrics such as effectiveness, efficiency and satisfaction at each turn for both systems. We observe that AESC performed better ($\mu = 31.16$, $\sigma = 14.5$) against baseline ($\mu = 24.33$, $\sigma = 16.83$), but had no significance $p = 0.052$ in terms of effectiveness. Satisfaction was much better ($\mu = 4.93$, $\sigma = 1.61$) versus baseline ($\mu = 4.18$, $\sigma = 1.92$) with $p = 0.036$. Time required to complete the task for AESC ($\mu = 4.1$, $\sigma = 1.74$) was more against baseline ($\mu = 3.25$, $\sigma = 1.59$) with statistical significance of $p = 0.027$. We can correlate this with the above interaction table 4.7, more the interaction, more the time taken.

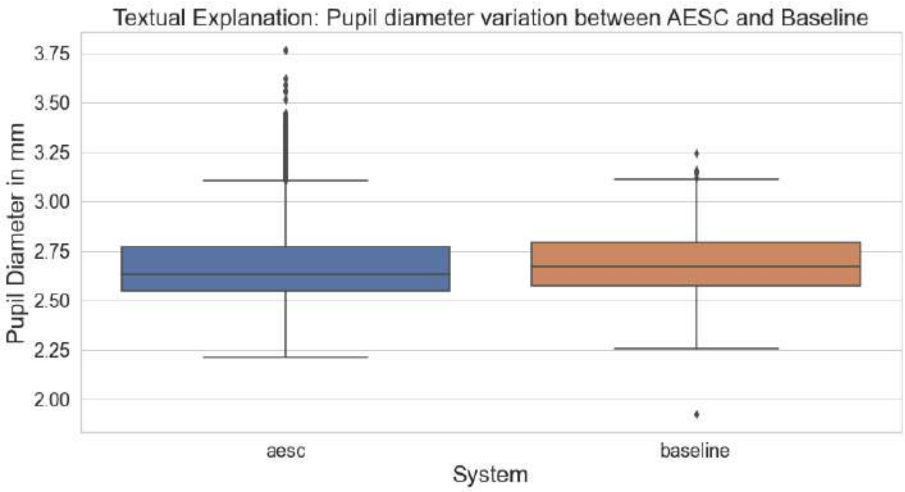
Table 4.8: RQ3: Comparison of usability metrics

Metric	AESC	Baseline
Effectiveness (in %)	**31.16 ± 14.5**	24.33± 16.83
Efficiency (in min)	4.1 ± 1.74	**3.25 ± 1.69**
Satisfaction	**4.93 ± 1.61**	4.18 ± 1.92

Cognitive load for baseline and AESC remains similar. However, AESC has more variations at higher end of pupil diameter indicating that, task load was varied among users. Examination of cognitive load across both systems show the same picture 4.8. AESC had slightly more pupil diameter of ($\mu = 2.7$, $\sigma = 0.24$) than baseline ($\mu = 2.68$, $\sigma = 0.17$) with no significance $p = 0.13$. This goes well with our hypothesis, that better explanations will have more interaction and lesser load.

Examination of gaze plots and heatmaps for AESC versus baseline system 4.9 show a similar picture. We see a bit more interaction with AESC as compared to baseline. This is also valid from interaction metrics 4.7. Observing no statistical significance of task effectiveness, We deduce that textual explanations ought to be refined more to be useful. Finally, we can conclude that personalization explanation was a mild success, it heavily suffered from generic texts, need to be filtered out to improve the explanation quality.

Figure 4.8: RQ3: Cognitive load variation between AESC and baseline



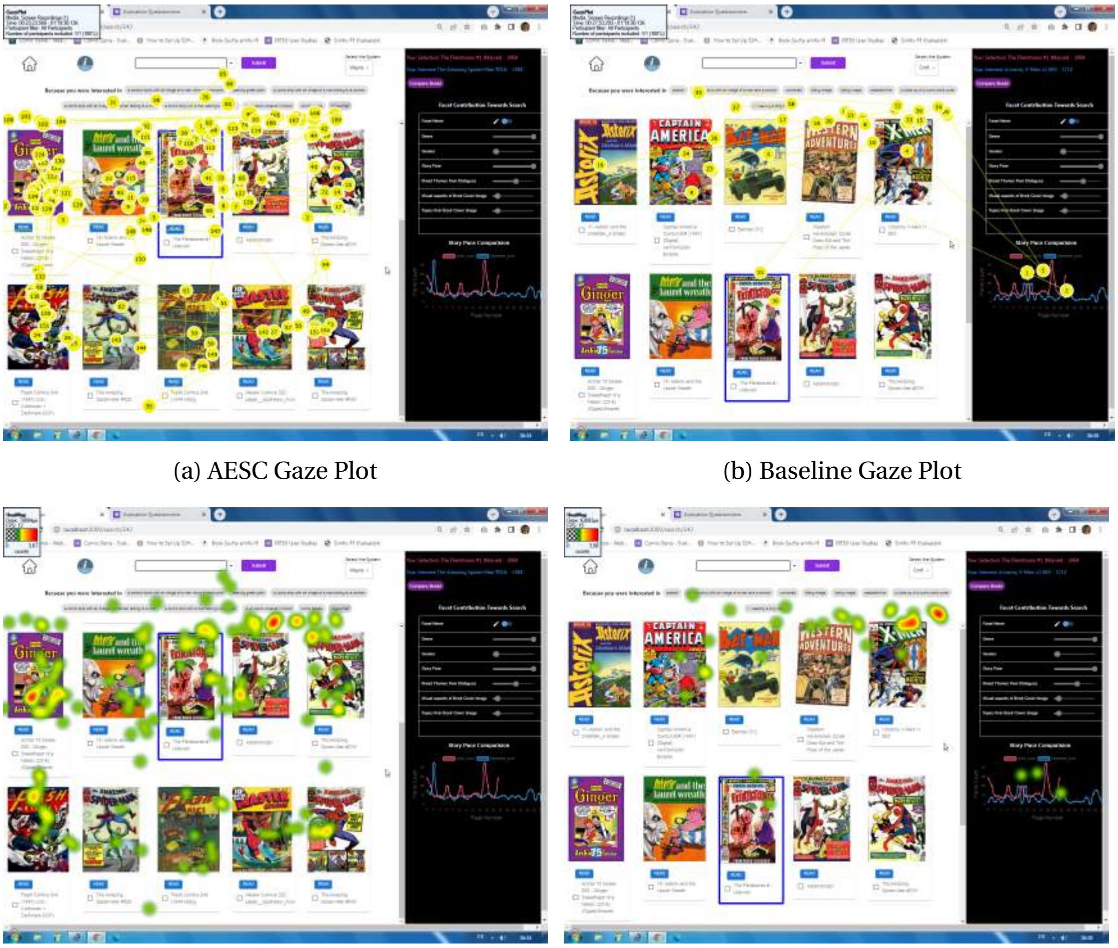
4.3.5 RQ 4: How does comparison table explanation help user discern between books from search results?

We attempt to answer, RQ4 through a task against both AESC and baseline. We pick a facet 'Story Pace' which is a bit obscure than genre, in order to evaluate the comparison explanation. We do not evaluate to comparison explanation against other explanation type, but only its usability in performing specific tasks. We ask user to perform task on both AESC and baseline. We record the interaction details and user's opinion.

Baseline System

Previous sections deliberated on using global and interaction explanation to illustrate system behaviour to user, but it doesn't empower user to compare books locally, within search results. A comparison table, that can compare multiple books on some of the facets could help users discern between books in search results. To evaluate usability of comparison table, we cannot compare system with comparison table against another system that doesn't have it, as it may naturally bias the user. We are not comparing the mode of explanation, but only whether the information it conveys is useful for the user. To evaluate comparison table we, create an identical UI with comparison table, but fill it with random information to create illusion of normalcy. Rest of the system remain identical to AESC.

Figure 4.9: RQ3: Gaze plot and heat map variations between AESC and baseline



(a) AESC Gaze Plot

(b) Baseline Gaze Plot

(c) AESC Heat Map

(d) Baseline Heat Map

Task

User interacts with the system for single turns. They are directly prodded to choose appropriate book and comparison explanation component, but are indirectly elicited about explanation information to achieve task goals. As task, User is asked to choose a lighter-read book than 'Justice League'. First, user chooses 'Justice League' book from landing page as query. They select appropriate books from evaluation task to compare against the query. Here, user has to self-understand the feature to use to achieve the task. Finally, they record books that are lighter-red than 'Justice League'. Failing to do so indicates, comparison table may not be best suited for this task. We also ask users to record their thoughts behind their choice, However this is non-mandatory. Task is repeated across both systems. We cross check users fast-paced books with our relevant fast-paced books list and calculate effectiveness 4.4

User assigns a score between 1-7 on likert scale for satisfaction about the task performance in both systems. Baseline system could have lesser relevant books than AESC, as it is not retrieved through facets.

Discussion

Table 4.9 provides captured metrics for count of hovered books and explanations checked for both systems. We deduce that, closed fact finding task made user select particular books from the search results. Hence the interaction difference between both the systems remain almost the same. However, users hovered more books from baseline ($\mu = 4.97$, $\sigma = 0.99$) against AESC ($\mu = 3.92$, $\sigma = 0.97$). From observing logs, users tend to hover more on baseline systems as it was showing different data than AESC. Only couple of users went to turn 2 by mistake and clicked on explanations, later they came back to previous page to complete the task. Hence baseline system had bit more interaction than AESC with statistical significance of $p = 0.032$

Table 4.9: RQ4: Interaction details for AESC vs Baseline

AESC		**Baseline**	
Hovered Books	**Explanations Checked**	**Hovered Books**	**Explanations Checked**
3.92 ± 0.97	**0.1 ± 0.01**	**4.97 ± 0.99**	0.1 ± 0.01

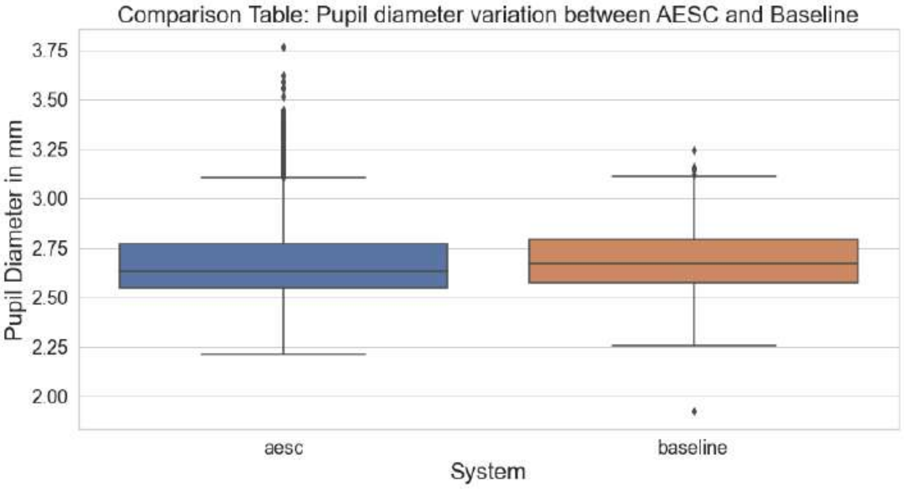
Table 4.10 provides captured usability metrics such as effectiveness, efficiency and satisfaction at each turn for both systems. We observe that AESC performed better ($\mu = 84$, $\sigma = 26.5$) against baseline ($\mu = 59$, $\sigma = 26.5$) with $p = 0.0288$ in terms of effectiveness. Satisfaction was also marginally better ($\mu = 5.87$, $\sigma = 1.21$) versus baseline ($\mu = 5.31$, $\sigma = 1.59$), but it was not statistically significant $p = 0.13622$. Time required to complete the task for AESC ($\mu = 4.11$, $\sigma = 1.84$) was more against baseline ($\mu = 3.25$, $\sigma = 1.59$) with statistical significance of $p = 0.027$. It was evident from table 4.9 and logs that, some users hovered multiple times on task related books in result set to figure out different data in baseline, hence the time of interaction was more than AESC. Users who did not cross check with visual explanation, tend to complete the task much quicker resulting in large standard deviation.

Increased cognitive load for baseline indicates more effort applied by users to solve the task. This is actually correlating with observed logs and interaction details. Examination of cognitive load across both turns for both systems show the same picture 4.10. AESC

Table 4.10: RQ4: Comparison of usability metrics

Metric	AESC	Baseline
Effectiveness (in %)	**84 ± 26.5**	59 ± 26.5
Efficiency (in min)	**2.91 ± 1.82**	5.01 ± 4.59
Satisfaction	**5.87 ± 1.21**	5.31 ± 1.59

Figure 4.10: RQ4: Cognitive load variation between AESC and baseline



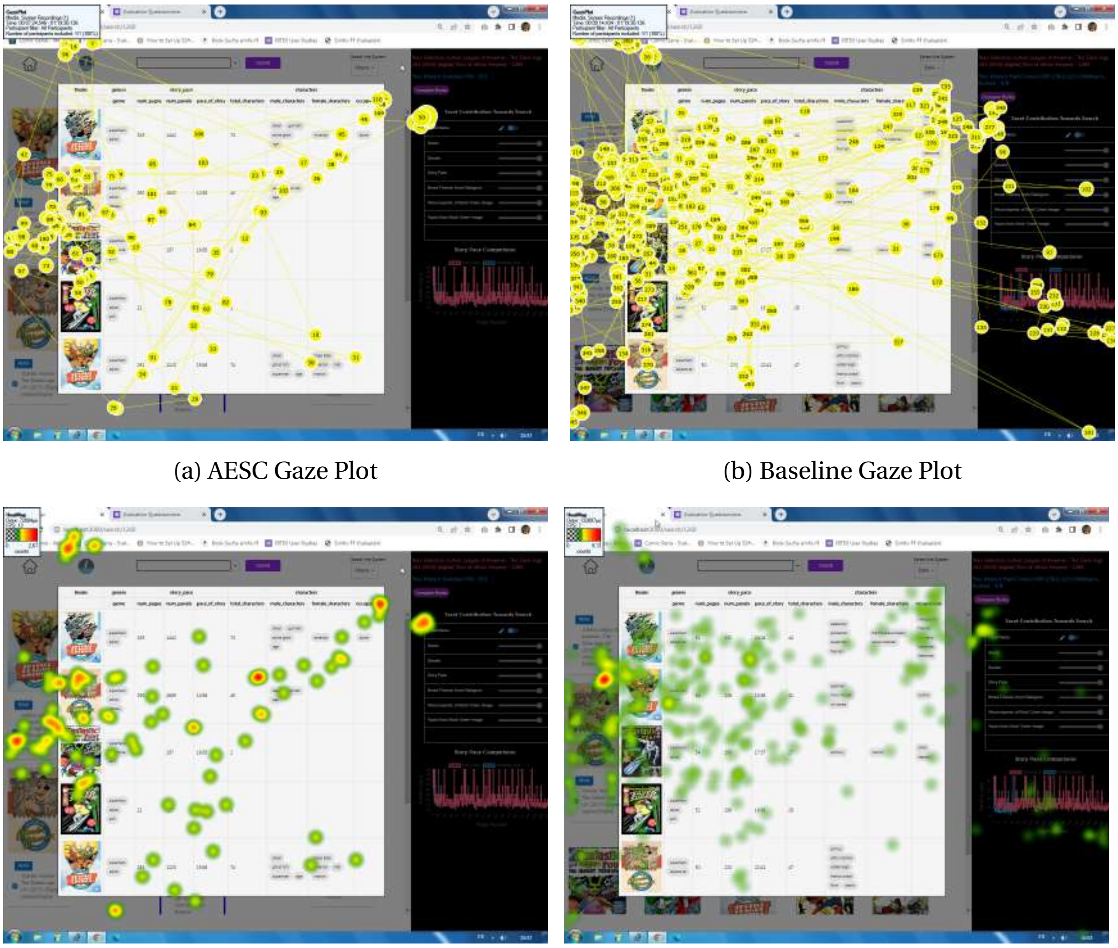
had lesser pupil diameter of ($\mu = 2.61$, $\sigma = 0.13$) than baseline ($\mu = 2.72$, $\sigma = 0.17$) with significance $p = 0.0001$. This goes well with our hypothesis, that AESC comparison table would have lesser load and interaction.

Examination of gaze plots and heatmaps for AESC versus baseline system 4.11 show the same picture. We see a bit more examination of comparison table, this may have increased cognitive load, but might not be the only factor. Finally, we can conclude that comparison table is a helpful component to compare books in SERP.

4.3.6 Supplementary Information

Users also recorded textual feedback about their though process against every task and also gave final feedback on deficiencies of the system. Users choose to hover only if they like the book cover. Genre was also mentioned as a main deciding factor to choose relevant books, unless task was to specifically ignore it like RQ2. Even for RQ4, users still used genre with story pace as shown in heatmaps. Users also thought that textual explanation that explain personalization is not reliable as it has more chaff than grain.

Figure 4.11: RQ4: Gaze plot and heat map variations between AESC and baseline



(a) AESC Gaze Plot

(b) Baseline Gaze Plot

(c) AESC Heat Map

(d) Baseline Heat Map

There were some really generic texts ‘a bunch of comic panels’, ‘high quality print’ and others which irritated the users, Hence the effectiveness went down for RQ3.

Some users suggested to include summary of book instead of topics, although small number of users felt topics are better and shorter. Visual explanation was appealing, but takes more effort to compare. Few users suggested to add character personality as facet. Few users wanted negative slider interaction, to get opposite books in the search results and escape the bubble. Few users advised to alter the slider from emphasizing to categorization, for example: increasing the gender slider more would gives more male oriented books and vice-versa rather than changing weights. Few users evaluation to be shortened or split up. Few users mentioned that they had difficulty in understanding interaction task RQ2, as the textual details were more.

Table 4.11: Preliminary evaluation against contextual variables such as gender and information retrieval background

Contextual Measure	Research Question	System	Interactive Precision	Effectiveness	Efficiency	Satisfaction
No Background	RQ1	AESC	71.06 ± 27.36	54.37 ± 18.69	2.75 ± 1.68	5.18 ± 1.51
		Baseline	57.28 ± 36.76	31.25 ± 21.75	1.69 ± 0.1	3.25 ± 1.98
	RQ2	AESC	73.36 ± 24.28	61.85 ± 24.8	7.09 ± 6.28	5.25 ± 1.08
		Baseline	59.56 ± 35.87	58.75 ± 22.04	3.2 ± 0.76	4.62 ± 1.45
	RQ3	AESC	-	5.06 ± 1.43	4.3 ± 1.89	1.43 ± 0.71
		Baseline	-	4.43 ± 1.83	3.38 ± 2.25	1.43 ± 0.933
	RQ4	AESC	-	6 ± 0.93	1.82 ± 1.95	1.68 ± 0.46
		Baseline	-	5.68 ± 1.26	6.69 ± 5.84	1.25 ± 0.43
Background	RQ1	AESC	83.04 ± 20.95	40.62 ± 30.1	2.19 ± 1.32	4.87 ± 1.45
		Baseline	46.87 ± 37.36	19.37 ± 17.48	1.47 ± 0.78	3.875 ± 2.02
	RQ2	AESC	77.84 ± 17.81	71.3 ± 24.9	6.93 ± 4.98	5.56 ± 1.93
		Baseline	63.93 ± 34.12	62.5 ± 20.9	3.12 ± 2.24	4.25 ± 1.98
	RQ3	AESC	-	4.81 ± 1.7	3.9 ± 1.89	2.31 ± 0.76
		Baseline	-	3.93 ± 1.92	3.12 ± 1.23	1.5 ± 1.06
	RQ4	AESC	-	5.75 ± 1.39	4 ± 1.69	1.68 ± 0.58
		Baseline	-	4.93 ± 1.75	3.33 ± 1.34	1.12 ± 0.6
Male	RQ1	AESC	75.73 ± 26.87	52.17 ± 24.66	2.21 ± 1.41	4.91 ± 1.58
		Baseline	54.34 ± 39.89	26.52 ± 21.18	1.39 ± 0.36	3.3 ± 2.11
	RQ2	AESC	78.67 ± 18.99	71.27 ± 35.58	5.79 ± 5.36	5.21 ± 1.31
		Baseline	62.44 ± 36.96	65.21 ± 23.19	1.9 ± 0.79	4.52 ± 1.63
	RQ3	AESC	-	4.81 ± 1.33	3.9 ± 1.81	1.72 ± 0.75
		Baseline	-	3.95 ± 1.94	3.07 ± 1.81	1.45 ± 0.98
	RQ4	AESC	-	5.7 ± 1.16	2.52 ± 1.66	1.68 ± 0.46
		Baseline	-	5.22 ± 1.75	5.71 ± 4.85	1.22 ± 0.51
Female	RQ1	AESC	77.96 ± 20.74	42 ± 30.1	2.73 ± 1.32	5.5 ± 1.2
		Baseline	51.66 ± 23.52	24.1 ± 18.54	1.77 ± 0.32	4.5 ± 1.74
	RQ2	AESC	65.97 ± 24.27	63 ± 25.31	8.23 ± 5.9	6 ± 1.26
		Baseline	63.97 ± 30.86	49 ± 17	4.42 ± 2.21	4.5 ± 1.91
	RQ3	AESC	-	5.2 ± 1.98	4.3 ± 1.67	2.2 ± 0.98
		Baseline	-	4.7 ± 1.67	3.43 ± 1.54	1.5 ± 1.02
	RQ4	AESC	-	6.1 ± 1.22	3.3 ± 1.98	1.7 ± 0.64
		Baseline	-	5.5 ± 1.02	5.44 ± 4.31	1.1 ± 0.53

A cursory look at gender based differences in evaluation would reveal that, female participants took more time to complete tasks. Co incidentally, they were more picky in selecting the books resulting in lower scores for effectiveness and precision, but had more satisfaction than male participants. From preliminary analysis of evaluation between participants with or without information retrieval background reveals that, time taken to complete all the tasks were less for participants with background, participants with information retrieval background had better precision and effectiveness in general. We can draw some conclusions that, our interaction component of search engine could be made more easier. Results are summarized in table 4.11

4.4 Summary of Evaluations

In this section, we summarize the results of the evaluations. We began for opting user study as a path for evaluating our search engine. This was due to non-availability of comic book retrieval dataset benchmarks. We collect contextual data related to demography, expertise and domain familiarity. Usability metric is also a main yardstick to evaluate AESC, provides a way to record users individual opinion and performance. We also measure goodness of system through IP. Finally, We complement the above measures with eye-tracker. We do not use eye-tracker to prove one system against the another, but to complement the above. We track changes in pupil diameter size to measure cognitive load for the task. We also analyze gaze plots of users to analyze their behaviour.

RQ1 : AESC had better search quality, thereby IP and effectiveness than baseline. However, it took more time in AESC, due to increased relevant books. Overall interaction with AESC is slightly better than baseline. IP is a higher than effectiveness as users selected fewer books that were more correct, increasing IP. Correlation between increased interaction with increased cognitive load is shaky.

RQ2 : As expected, AESC was able to understand user's explicit request and provided better search results increasing IP, effectiveness and satisfaction. IP is a higher than effectiveness as users selected fewer books were more correct increasing IP. Tasks on AESC took more time to solve than baseline, due to increased relevant books. Users were fatigued while fiddling with sliders, maybe better UI component is needed to convey their interests. This is different than hypothesis.

RQ3 : Textual explanations that can explain by personalization did not succeed as expected. Even though it was slightly better than baseline. Some of the reasons include generic textual description like 1a comic panel', 'blu-ray resolution' and others. Probably, textual captions should be filtered first and explained later. Interaction with AESC was slightly better than baseline.

RQ4 : Overall interaction with AESC is less than baseline with significance. This was due to the fact that some users cross-verified visual explanation with table results. Since baseline contradicted results between table and visual explanation, their interaction was more and they took more time as well. Users were more effective, efficient and satisfied with AESC significantly than baseline. Cognitive load conforms with mentioned hypothesis.

We summarize the results for different metrics in following table 4.12. Metrics that have achieved significance will be italicised. An overall trend signified that AESC was more effective than baselines and satisfied user in their information need. However, tasks on AESC took more time than baseline. This is task dependent, but a general trend is that more relevant artifacts, more the time taken. Interaction is a double-edged sword and inconsistent with hypothesis.

Table 4.12: Summary of evaluation with metrics. Y signifies better performance than N. Performance difference with statistical significance is *italicised*.

	Interactive Precision		Effectiveness		Efficiency		Satisfaction	
	AESC	Baseline	AESC	Baseline	AESC	Baseline	AESC	Baseline
RQ1	***Y***	N	***Y***	N	N	***N***	***Y***	N
RQ2	***Y***	N	***Y***	N	N	***N***	***Y***	N
RQ3	-	-	Y	N	N	***Y***	***Y***	N
RQ4	-	-	***Y***	N	***Y***	N	***Y***	N

5
Conclusion and Future Work

5.1 Conclusion

We proposed an adaptive comic information retrieval system that considered book content, took into account search context, and explained search results.

First, we highlight key facets used by users to sift across comic books. We extract visual facets like color, texture, edges, and book covers. Textual facets including genre, gender, book cover details, and coarse categories. Furthermore, we derive temporal facet such as total information in the book and panels per page. We create a hybrid dataset by mixing comics from the DCM COMICS dataset with renowned comics from the similar age and culture.

Not every facet can be the same and makes sense to the user. Hence we create a two-tiered feature space with lower level consisting of color, texture, edges and text. Top-200 filtered results from lower level features were re-ranked by higher level domain based facets.

Not every user expects same results for a given query, there are always some context surrounding the search. Hence we assume users interest and non-interest based on hovering activity. A simple GD based model picks up these activities and hones the facet weights accordingly using TBA. Users can explicitly emphasize or ignore their preferences to the system through facet weights.

Global explanations are given in the form of facet weights. We explain personalization by through book cover's textual description. Local explanations such as story pace using visual chart, and a textual comparison table for multiple books. Topics extracted from comics are presented to provide additional information on hover.

Finally, we conducted a user study to assess our search engine. Due to the absence of benchmarks, we collected both contextual data collection and observation of user behavior by employing usability metrics and eye-tracking technology. Overall results suggest that although AESC outperforms baseline in interaction, effectiveness and user satisfaction, there was room for improvement in terms of efficiency. AESC provided much better search performance than pseudo random retriever. Users were able to convey their needs better with AESC, albeit some fatigue. Furthermore, users were able to compare between different books in SERP effectively with comparison table. On the contrary, explaining personalization suffered from generic description rather than pinpointed one. Across all the tasks, AESC took more time than baselines. Most of the users mentioned that they used book cover and genre to decide relevant books. Users suggested, summary of the book would be better for deciding it's relevance.

5.2 Future Work

Our first limitation is the non-availability of any benchmarks for comics book retrieval. Availability of benchmarks makes it easier to evaluate any new retrieval algorithm, facet or explanation. We alleviate this through initial evaluation of manual-dataset 3.2. User study evaluation could be difficult to reproduce with a different distribution of users. Sometimes it is difficult to elicit from same users as well. Creating a robust dataset along with benchmarks verified by domain experts would be the really important.

Facets to be extracted are decided based on user's choice for its utilization. Comic books are a rich visual medium than textual. Visuals carry the narrative forward and texts only support them. We used PCA dimensionality reduction rather than T-SNE due to computation constraints. Complex and rich visual facets like page layout describes multiple elements about comics such as pacing, artistic style. Background and foreground is an important artistic medium. Emotions could be expressed as onomatopoeia. Transition of panels signifies narrative shift. We completely forego these facets due to complexity and dataset issues. Semantic facet based retrieval only served as a demonstrator rather than solid model against existing powerful retrieval system. Better benchmark could be created to further test the retriever.

Adaptation according to surrounding context behind users search is an important aspect of the thesis. First problem is that we assume that user's hover equates to relevance of the book, this is debatable. Secondly, we assume that search context won't change rapidly during search. Lastly, we are only considering previous turn interaction and omit complete history. We used a simplistic GD model that uses TBA to optimize facet weights. GD has it's own problems, getting stuck in a local optima, filter bubbles, difficult to handle large or low gradients and many more. We can utilize history of the user for more then previous turn to better optimize. A better adaptation method based on reinforcement learning or a feature importance based model or a improved GD based model could be used. Furthermore, users may not want to provide feedback through weights, instead they could provide some text that describes their context and system could learn to set the weights on its own.

Personalization explanation had lot off generic prompts making it unusable. Furthermore, these explanations can be linked to other previous turns in the history creating a chain of explanations. Visual explanation about story pace was attractive, but less people found it useful as using a single number or tag from comparison table felt easier. Hence this can be replaced with tags like ‘fast-paced’ and ‘slow-paced’. Sliders can be enhanced to accommodate negative weights to provide opposite search results. We can use other techniques related to feature importance as weights of facets instead of GD. LIME can be used to explain locally about keywords responsible for retrieval. Summary of the book would be helpful for understanding the content better.

Evaluation was reported as lengthy process by many users. It can be shortened and randomized appropriately. Time was consumed more on RQ2 1.2, hence text based interaction could be used. We did not evaluate UI components, like comparison table versus visual explanation or personalization explanation versus facet weights to know their relative effectiveness. Evaluating performance of retriever on a benchmark and including more metrics related to interaction such as screen dwell time, search bar used to discontinue search, re-retrieval with user provided weights and others could be done. An open-ended task of continuously exploring AESC versus baseline for more than 10 turns could solidify the evaluation.

A
Appendix

A.1 Evaluation Form and Search Engine Details

Below format is standardized using Liu et al. [165].

1. Research Focus: domain based facets, implicit feedback, interactivity, explainability; *Independent Variable*: Facets, Facet Weights, Textual Explanations, Comparison Table Explanation; *Dependent Variable*: search performance, task efficiency, satisfaction, usability, think aloud experience;

2. Participant *Recruitment: Doodle and OVGU university emailing system*; *Controlled lab*: Yes; *Sample Size*: 32 participants; *Gender Composition:* 22 male, 10 female; *Participant Background* students and researchers; *Age*: 24 between 18 and 30 years, 8 between 30 and 40 years; *Education Background/Level*: graduate; *Language Used in Study*: English; *Regular Incentives*: N.A; *Extra Incentives/Bonus*: Refreshments and Beverage; *Length of Study*: approximate 80 minutes.

3. Tasks

Task Source: N.A; *Search Task Type*: controlled exploratory, controlled interactive, fact-finding amorphous, fact-finding specific; *Search Task Topic* 4 initial books; *Number of Tasks*: 4; *Number of tasks/Person*: 4 tasks; *Time Length/Task*: 15 min per task; *Work Task Type*: problematic situation/context provided in task descriptions; *Did Work Task*: Yes; *Answer Search Task*: Yes; *Evaluation Task*: post-search rating satisfaction

Study Procedure and Experimental Design *Task/Session Feature Controlled*: same 4 types for every individual users; *Task Rotation*: Graeco-Latin-square rotation of task and system order; *Pilot Study*: N.A; *Pre-study Training*: before 4 formal tasks, each participant was briefed on the system and freely explored the system for 5-10 minutes; *Actual Task Completion Time* about 15 min per task; *Quality Control/Data Filtering*

Criteria: English speakers and incomplete evaluation, 1 user excluded; [Experimental Design] within-subjects design

4. System Features

Study Interface Element Varied: No; *Other System/Context Feature Varied*: Yes; *Study Apparatus* desktop computer, with and without eye-tracker ; *Search Collection/Corpus*: custom comic book dataset based on COMICS; *Ranking Algorithm*: Two-Tiered, Reweighing, Cosine Similarity; *Non-traditional IR System Assistance Tool*: N.A.

5. Behavioral and Search Experience Measures

Search Behavior Measures: hovered books; explanations clicked; **information available before users clicking on the result**: local and global explanations ; **Instrument for Collecting Search Behavioral Data**: logged by an experimental system; **Relevance Judgment**: post-search relevance judgment; **Instrument for Collecting User Judgment**: Questionnaire from Google Form; **Search and System Performance Measures**: Interactive Precision, Effectiveness, Task Efficiency ; **Neuro-physiological Measures**: Pupil Diameter; **Instruments for Capturing Neuro-physiological Measures**: Tobii Eye-Tracker; **Data Analysis Method**: Wilcoxon-Signed Two Tailed Test; **Qualitative Analysis**: Recording of eye-tracker; **Level of Analysis**: task level; turn level; **Task-independent Measures**: pre-search survey about gender, age, background, domain experiences ; **Task/Session Perception Measures**: N.A; **post-search survey**: N.A; **Search Experience and System Evaluation Measures**: post-search survey: satisfaction, pain-points.

6. Data Analysis and Results

Statistical Test Assumption Check Reported: Gender or background not considered for evaluation; **Results**: statistical significance about AESC performing better than baseline for search quality, interaction and comparison table | no significance for personalization explanation usefulness | complementing recorded measure with neuro-physical measure like cognitive load, but not used to make any claims about the system;

7. Evaluation Form Link URL

evaluation form url : `https://forms.gle/c9JyfmTfCs8Ej72V9`

8. Search Engine

Code for comic book search engine can be accessed in github repository at `https://github.com/surajsrivathsa/thesis_deployment.git`

B
Abbreviations and Notations

Acronym	Meaning
AESC	Adaptive and Explainable Search for Comics System
ANN	Artificial Neural Network
BPTT	Backpropagation Through Time
CNN	Convolutiona Neural Networks
DL	Deep Learning
DNN	Deep Neural Network
EL2N	Error L2 Norm
EL1N	Error L1 Norm
GD	Gradient Descent
IP	Interactive Precision
LSTM	Long Short Term Memory
LOD	Linked Open Data
MBSE	Metadata based Search Engine
ML	Machine Learning
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
RNN	Recurrent Neural Network
SERP	Search Engine Results Page
TBA	Triplet based Adaptivity

C

List of Figures

D
List of Tables

E

Bibliography

[1] T. N. Le, M. M. Luqman, J.-C. Burie, and J.-M. Ogier, “Retrieval of comic book images using context relevance information,” *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, 2016. pages

[2] N.-V. Nguyen, C. Rigaud, and J.-C. Burie, “What do we expect from comic panel extraction?,” *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 1, pp. 44–49, 2019. pages

[3] Y. Daiku, M. Iwata, O. Augereau, and K. Kise, “Comics story representation system based on genre,” *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 257–262, 2018. pages

[4] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, “Sketch-based manga retrieval using manga109 dataset,” *Multimedia Tools and Applications*, vol. 76, pp. 21811–21838, 2015. pages

[5] A. Foster and N. Ford, “Serendipity and information seeking: an empirical study,” *J. Documentation*, vol. 59, pp. 321–340, 2003. pages

[6] N.-V. Nguyen, C. Rigaud, and J.-C. Burie, “Digital comics image indexing based on deep learning,” *J. Imaging*, vol. 4, p. 89, 2018. pages

[7] D. S. Lenadora, R. Ranathunge, C. Samarawickrama, Y. D. Silva, I. Perera, Anuradha, and A. Welivita, “Extraction of semantic content and styles in comic books,” *International Journal on Advances in Ict for Emerging Regions (icter)*, vol. 13, 2020. pages

[8] S. Stober and A. Nürnberger, “Adaptive music retrieval–a state of the art,” *Multimedia Tools and Applications*, vol. 65, pp. 467–494, 2013. pages

[9] A. Anand, L. Lyu, M. Idahl, Y. Wang, J. Wallat, and Z. Zhang, “Explainable information retrieval: A survey,” *ArXiv*, vol. abs/2211.02405, 2022. pages

[10] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Inf. Fusion*, vol. 58, p. 82–115, jun 2020. pages

[11] M. Iyyer, V. Manjunatha, A. Guha, Y. Vyas, J. L. Boyd-Graber, H. Daumé, and L. S. Davis, “The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6478–6487, 2016. pages

[12] E. Hoffer and N. Ailon, “Deep metric learning using triplet network,” in *International Workshop on Similarity-Based Pattern Recognition*, 2014. pages

[13] S. Stober, “Adaptive methods for user-centered organization of music collections,” 2011. pages

[14] T. Mitchell, *Machine Learning*. McGraw-Hill International Editions, McGraw-Hill, 1997. pages

[15] C. M. Bishop and N. M. Nasrabadi, “Pattern recognition and machine learning,” *J. Electronic Imaging*, vol. 16, p. 049901, 2006. pages

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`. pages

[17] Müller, Meinard, *Fundamentals of Music Processing*. Cham: Springer International Publishing, 01 2021. pages

[18] F. N. Catbas, T. L. Kijewski-Correa, and A. E. Aktan, “Structural identification of constructed systems : approaches, methods, and technologies for effective practice of st-id,” 2013. pages

[19] Q. Wang, Y. Ma, K. Zhao, and Y. jie Tian, “A comprehensive survey of loss functions in machine learning,” *Annals of Data Science*, pp. 1–26, 2020. pages

[20] M. Schultz and T. Joachims, “Learning a distance metric from relative comparisons,” in *NIPS*, 2003. pages

[21] A. Gunawardana and G. Shani, “A survey of accuracy evaluation metrics of recommendation tasks,” *J. Mach. Learn. Res.*, vol. 10, pp. 2935–2962, 2009. pages

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017. pages

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations, ICLR 2015*, (San Diego, United States), jan 2015. pages

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020. pages

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. pages

[26] R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval.* ACM Press / Addison-Wesley, 1999. pages

[27] M. A. Hearst, "Design recommendations for hierarchical faceted search interfaces," in *Proc. SIGIR 2006, Workshop on Faceted Search*, pp. 26–30, August 2006. pages

[28] J. Koenemann and N. J. Belkin, "A case for interaction: A study of interactive information retrieval behavior and effectiveness," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 205–212, 1996. pages

[29] S. Pauws and B. Eggen, "Realization and user evaluation of an automatic playlist generator," *Journal of New Music Research*, vol. 32, pp. 179 – 192, 2003. pages

[30] A. Nürnberger and M. Detyniecki, "Weighted self-organizing maps: Incorporating user feedback," in *International Conference on Artificial Neural Networks*, 2003. pages

[31] T. Ruotsalo, J. Peltonen, M. J. A. Eugster, D. Glowacka, A. Reijonen, G. Jacucci, P. Myllymäki, and S. Kaski, "Scinet: Interactive intent modeling for information discovery," *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015. pages

[32] T. Ruotsalo, G. Jacucci, and S. Kaski, "Interactive faceted query suggestion for exploratory search: Whole-session effectiveness and interaction engagement," *Journal of the Association for Information Science and Technology*, vol. 71, pp. 742 – 756, 2019. pages

[33] A. Medlar, J. Li, and D. Glowacka, "Query suggestions as summarization in exploratory search," *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 2021. pages

[34] G. Marchionini, "Exploratory search: From finding to understanding," *Commun. ACM*, vol. 49, p. 41–46, apr 2006. pages

[35] V. G. Dimitrova, L. M. S. Lau, D. Thakker, F. Yang-Turner, and D. Despotakis, "Exploring exploratory search: a user study with linked semantic data," in *IESD '13*, 2013. pages

[36] N. Marie, "Linked data based exploratory search," 2014. pages

[37] T.-N. Nguyen, D. Dinh, and T.-D. Cao, "Empowering exploratory search on linked movie open data with semantic technologies," *Proceedings of the 6th International Symposium on Information and Communication Technology*, 2015. pages

[38] K. Athukorala, D. Glowacka, G. Jacucci, A. Oulasvirta, and J. Vreeken, "Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks," *Journal of the Association for Information Science and Technology*, vol. 67, 2016. pages

[39] T. Low, C. Hentschel, S. Polley, A. Das, H. Sack, A. Nürnberger, and S. Stober, "The ismir explorer - a visual interface for exploring 20 years of ismir publications," in *International Society for Music Information Retrieval Conference*, 2019. pages

[40] Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.-S. Chua, and X.-S. Hua, "Visual query suggestion: Towards capturing user intent in internet image search," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, aug 2010. pages

[41] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, pp. 262–282, 2007. pages

[42] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. pages

[43] J. J. Rocchio, "Relevance feedback in information retrieval," 1971. pages

[44] I. Ruthven and M. Lalmas, “A survey on the use of relevance feedback for information access systems,” *The Knowledge Engineering Review*, vol. 18, pp. 95 – 145, 2003. pages

[45] J. Langford and T. Zhang, “The epoch-greedy algorithm for multi-armed bandits with side information,” in *NIPS*, 2007. pages

[46] S. Chaudhuri and A. Tewari, “Online learning to rank with top-k feedback,” *J. Mach. Learn. Res.*, vol. 18, pp. 103:1–103:50, 2016. pages

[47] S. Li, T. Lattimore, and C. Szepesvari, “Online learning to rank with features,” *ArXiv*, vol. abs/1810.02567, 2018. pages

[48] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay, “Accurately interpreting clickthrough data as implicit feedback,” *SIGIR Forum*, vol. 51, pp. 4–11, 2005. pages

[49] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay, “Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search,” *ACM Trans. Inf. Syst.*, vol. 25, p. 7, 2007. pages

[50] A. Chuklin, P. Serdyukov, and M. de Rijke, “Click model-based information retrieval metrics,” *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013. pages

[51] C. Buckley and G. Salton, “Optimization of relevance feedback weights,” in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995. pages

[52] Y. Rui, T. S. Huang, M. Ortega-Binderberger, and S. Mehrotra, “Relevance feedback: a power tool for interactive content-based image retrieval,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 644–655, 1998. pages

[53] B. Liu, Y. Cao, M. Long, J. Wang, and J. Wang, “Deep triplet quantization,” *Proceedings of the 26th ACM international conference on Multimedia*, 2018. pages

[54] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision*, 2014. pages

[55] A. Krizhevsky, “Learning multiple layers of features from tiny images,” tech. rep., 2009. pages

[56] D. R. Chittajallu, B. Dong, P. Tunison, R. Collins, K. O. Wells, J. W. Fleshman, G. Sankaranarayanan, S. D. Schwaitzberg, L. A. Cavuoto, and A. Enquobahrie, “Xai-cbir: Explainable ai system for content based retrieval of video frames from minimally invasive surgery videos,” *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 66–69, 2019. pages

[57] D. M. J. M. M. D. M. N. P. Andru Twinanda, Sherif Shehata, “Endonet: A deep architecture for recognition tasks on laparoscopic videos,” *IEEE Transactions on Medical Imaging*, vol. 36, 02 2016. pages

[58] X. Yang, H. Qi, M. Li, and A. Hauptmann, “From a glance to "gotcha": Interactive facial image retrieval with progressive relevance feedback,” *ArXiv*, vol. abs/2007.15683, 2020. pages

[59] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild.,” in *ICCV*, pp. 3730–3738, IEEE Computer Society, 2015. pages

[60] L. Wang, X. Qian, Y. Zhang, J. Shen, and X. Cao, “Enhancing sketch-based image retrieval by cnn semantic re-ranking,” *IEEE Transactions on Cybernetics*, vol. 50, pp. 3330–3342, 2020. pages

[61] Y. Jia, H. Wang, S. Guo, and H. Wang, “Pairrank: Online pairwise learning to rank by divide-and-conquer,” *Proceedings of the Web Conference 2021*, 2021. pages

[62] A. C. Scott, W. J. Clancey, R. Davis, and E. H. Shortliffe, “Explanation capabilities of production-based consultation systems,” *American Journal of Computational Linguistics*, pp. 1–50, Feb. 1977. Microfiche 62. pages

[63] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, pp. 484–489, 2016. pages

[64] N. Burkart and M. F. Huber, “A survey on the explainability of supervised machine learning,” *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2020. pages

[65] S. M. Lundberg, G. G. Erion, and S.-I. Lee, “Consistent individualized feature attribution for tree ensembles,” *ArXiv*, vol. abs/1802.03888, 2018. pages

[66] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International Conference on Machine Learning*, 2017. pages

[67] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. pages

[68] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, pp. 1–12, 2021. pages

[69] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," *ArXiv*, vol. abs/1909.03012, 2019. pages

[70] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, "Understanding the behaviors of bert in ranking," *ArXiv*, vol. abs/1904.07531, 2019. pages

[71] J. Singh and A. Anand, "Exs: Explainable search using local model agnostic interpretability," *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2018. pages

[72] S. Polley, A. Janki, M. Thiel, J. Hoebel-Mueller, and A. Nürnberger, "Exdocs: Evidence based explainable document search," 2021. pages

[73] S. Polley, S. Mondal, V. S. Mannam, K. Kumar, S. Patra, and A. Nürnberger, "X-vision: Explainable image retrieval by re-ranking in semantic space," *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022. pages

[74] S. Polley, R. R. Koparde, A. B. Gowri, M. Perera, and A. Nürnberger, "Towards trustworthiness in the context of explainable search," *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. pages

[75] J. Singh and A. Anand, "Posthoc interpretability of learning to rank models using secondary training data," *ArXiv*, vol. abs/1806.11330, 2018. pages

[76] J. Singh, M. Khosla, W. Zhenye, and A. Anand, "Extracting per query valid explanations for blackbox learning-to-rank models," *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, 2021. pages

[77] A. Purpura, K. Buchner, G. Silvello, and G. A. Susto, "Neural feature selection for learning to rank," in *European Conference on Information Retrieval*, 2021. pages

[78] Z. T. Fernando, J. Singh, and A. Anand, "A study on the interpretability of neural retrieval models using deepshap," *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019. pages

[79] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *ArXiv*, vol. abs/1705.07874, 2017. pages

[80] O. Khattab and M. A. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. pages

[81] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and A. Veneri, "Ilmart: Interpretable ranking with constrained lambdamart," *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022. pages

[82] C. J. C. Burges, "From ranknet to lambdarank to lambdamart: An overview," 2010. pages

[83] M. A. Hearst, "User interfaces for search by marti a . hearst," 2009. pages

[84] T. Low, "Towards combining search and exploration: escaping the filter bubble through map-based exploration / thomas low ; gutachter: Andreas nürnberger," 2022. pages

[85] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," *Found. Trends Inf. Retr.*, vol. 3, pp. 1–224, 2009. pages

[86] K. Mikkonen, "The narratology of comic art," 2017. pages

[87] N. Cohn, "Navigating comics: An empirical and theoretical approach to strategies of reading comic page layouts," *Frontiers in Psychology*, vol. 4, 2013. pages

[88] N. K. Pratha, N. Avunjian, and N. Cohn, "Pow, punch, pika, and chu: The structure of sound effects in genres of american comics and japanese manga," *Multimodal Communication*, vol. 5, pp. 109 – 93, 2016. pages

[89] N. V. Nguyen, C. Rigaud, and J. C. Burie, “Digital comics image indexing based on deep learning,” *Journal of Imaging*, vol. 4, 7 2018. pages

[90] J. Guo, N. Xu, X. Chen, Y. Shi, K. Xu, and A. Alwan, “Filter Sampling and Combination CNN (FSC-CNN): A Compact CNN Model for Small-footprint ASR Acoustic Modeling Using Raw Waveforms,” in *INTERSPEECH Conference*, (Hyderabad , India), International Speech Communication Association (ISCA), 2018. pages

[91] Y. In, T. Oie, M. Higuchi, S. Kawasaki, A. Koike, and H. Murakami, “Fast frame decomposition and sorting by contour tracing for mobile phone comic images,” 2010. pages

[92] L. Li, Y. Wang, Z. Tang, and L. Gao, “Automatic comic page segmentation based on polygon detection,” *Multimedia Tools and Applications*, vol. 69, pp. 171–197, 2012. pages

[93] A. K. N. Ho, J.-C. Burie, and J.-M. Ogier, “Panel and speech balloon extraction from comic books,” *2012 10th IAPR International Workshop on Document Analysis Systems*, pp. 424–428, 2012. pages

[94] R. B. Girshick, “Fast r-cnn,” *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015. pages

[95] J. Laubrock and D. Dubray, “Multi-class semantic segmentation of comics: A u-net based approach,” 07 2019. pages

[96] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *ArXiv*, vol. abs/1505.04597, 2015. pages

[97] A. Dunst, R. Hartel, and J. Laubrock, “The graphic narrative corpus (gnc): Design, annotation, and analysis for the digital humanities,” *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 03, pp. 15–20, 2017. pages

[98] S. M. Yusof, Z. M. Lazim, K. Salehuddin, and M. M. Shahimin, “Graphic novels: Understanding how fifth graders read literary texts through eye movements analysis,” *Kritika Kultura*, 2019. pages

[99] J. Wildfeuer, I. V. der sluis, G. Redeker, and N. van der Velden, “No laughing matter!? analyzing the page layout of instruction comics,” *Journal of Graphic Novels and Comics*, vol. 14, pp. 186 – 207, 2022. pages

[100] W. H. Bares, “Panel beat: Layout and timing of comic panels,” in *International Symposium on Smart Graphics*, 2008. pages

[101] K. Pederson and N. Cohn, “The changing pages of comics : Page layouts across eight decades of american superhero comics,” *Studies in Comics*, vol. 7, pp. 7–28, 2016. pages

[102] W. Sun, J.-C. Burie, J.-M. Ogier, and K. Kise, “Specific comic character detection using local feature matching,” *2013 12th International Conference on Document Analysis and Recognition*, pp. 275–279, 2013. pages

[103] G. LoweDavid, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, 2004. pages

[104] X. Qin, Y. Zhou, Y. Li, S. Wang, Y. Wang, and Z. Tang, “Progressive deep feature learning for manga character recognition via unlabeled training data,” *Proceedings of the ACM Turing Celebration Conference - China*, 2019. pages

[105] W.-T. Chu and W.-W. Li, “Manga facenet: Face detection in manga based on deep neural network,” *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 2017. pages

[106] N.-V. Nguyen, C. Rigaud, and J.-C. Burie, “Comic mtl: optimized multi-task learning for comic book image analysis,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, pp. 265 – 284, 2019. pages

[107] C. Forceville, T. Veale, and K. Feyaerts, “Balloonics: the visuals of balloons in comics,” 2010. pages

[108] J. M. C. Correia and A. J. P. Gomes, “Balloon extraction from complex comic books using edge detection and histogram scoring,” *Multimedia Tools and Applications*, vol. 75, pp. 11367–11390, 2016. pages

[109] C. Rigaud, J.-C. Burie, and J.-M. Ogier, “Text-independent speech balloon segmentation for comics and manga,” in *IAPR International Workshop on Graphics Recognition*, 2015. pages

[110] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, “Mask r-cnn,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 386–397, 2017. pages

[111] D. Dubray and J. Laubrock, “Deep cnn-based speech balloon detection and segmentation for comic books,” *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1237–1243, 2019. pages

[112] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. pages

[113] A. Dutta, S. Biswas, and A. K. Das, “Cnn-based segmentation of speech balloons and narrative text boxes from comic book page images,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 24, pp. 49 – 62, 2021. pages

[114] S. McCloud, *Understanding Comics: The Invisible Art.* Tundra Publishing Ltd, 1993. pages

[115] J. Baek, Y. Matsui, and K. Aizawa, “Coo: Comic onomatopoeia dataset for recognizing arbitrary or truncated texts,” in *European Conference on Computer Vision*, 2022. pages

[116] O. Vinyals, M. Fortunato, and N. Jaitly, “Pointer networks,” in *NIPS*, 2015. pages

[117] O. Rohan, R. Sasamoto, and S. O’Brien, “Onomatopoeia: A relevance-based eye-tracking study of digital manga,” *Journal of Pragmatics*, 2021. pages

[118] R. Takizawa and S. Hirai, “Synthesis of explosion sounds from utterance voice of onomatopoeia using transformer,” *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023. pages

[119] T. Okada, F. Toriumi, and M. Sakamoto, “A study on emotional analysis focusing on onomatopoeia used on sns for the covid-19,” *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2021. pages

[120] M. S. K. Devi, S. Fathima, and R. Baskaran, “Cbcs - comic book cover synopsis: Generating synopsis of a comic book with unsupervised abstractive dialogue,” *Procedia Computer Science*, vol. 172, pp. 701–708, 2020. pages

[121] W. Zheng and M. Jin, “A review on authorship attribution in text mining,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 15, 2023. pages

[122] C. Johnson, E. Hendriks, I. J. Berezhnoy, E. Brevdo, S. M. Hughes, I. Daubechies, J. Li, E. M. Postma, and J. Wang, “Image processing for artist identification,” *IEEE Signal Processing Magazine*, vol. 25, 2008. pages

[123] W. Sun and K. Kise, “Similar partial copy recognition for line drawings using concentric multi-region histograms of oriented gradients,” in *IAPR International Workshop on Machine Vision Applications*, 2011. pages

[124] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 886–893 vol. 1, 2005. pages

[125] A. Dunst and R. Hartel, “Automated genre and author distinction in comics: Towards a stylometry for visual narrative,” in *International Conference on Digital Health*, 2018. pages

[126] J. Laubrock and D. Dubray, “Cnn-based classification of illustrator style in graphic novels: Which features contribute most?,” in *Conference on Multimedia Modeling*, 2018. pages

[127] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2015. pages

[128] A. Dunst and R. Hartel, “Quantifying complexity in multimodal media: Alanmoore and the “density” of the graphic novel,” 2019. pages

[129] H. Yang, M. Kashimura, N. Onda, and S. Ozawa, “Extraction of bibliography information based on image of book cover,” *Proceedings 10th International Conference on Image Analysis and Processing*, pp. 921–926, 1999. pages

[130] B. K. Iwana and S. Uchida, “Judging a book by its cover,” *ArXiv*, vol. abs/1610.09204, 2016. pages

[131] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, pp. 84 – 90, 2012. pages

[132] G. R. Biradar, R. Jm, A. Varier, and M. Sudhir, “Classification of book genres using book cover and title,” *2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT)*, pp. 72–723, 2019. pages

[133] Y. Daiku, O. Augereau, M. Iwata, and K. Kise, “Comic story analysis based on genre classification,” *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 03, pp. 60–65, 2017. pages

[134] K. Sivaraman and G. Somappa, “Moviescope: Movie trailer classification using deep neural networks,” 2017. pages

[135] R. Bucher, “Classification of fiction genres text classification of fiction texts from project gutenberg,” 2019. pages

[136] G. Barney and K. Kaya, “Predicting genre from movie posters,” 2019. pages

[137] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *ACM Trans. Interact. Intell. Syst.*, vol. 5, pp. 19:1–19:19, 2016. pages

[138] C. Xu, X. Xu, N. Zhao, W. Cai, H. Zhang, C. Li, and X. Liu, “Panel-page-aware comic genre understanding,” *IEEE Transactions on Image Processing*, vol. 32, pp. 2636–2648, 2023. pages

[139] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. pages

[140] A. Dey, C. Radhakrishna, N. N. Lima, S. B. Shashidhar, S. Polley, M. Thiel, and A. Nürnberger, “Evaluating reliability in explainable search,” *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pp. 1–4, 2021. pages

[141] S. Polley, S. Ghosh, M. Thiel, M. Kotzyba, and A. Nürnberger, “Simfic: An explainable book search companion,” *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pp. 1–6, 2020. pages

[142] S. K. Kondreddi, P. Triantafillou, and G. Weikum, “Combining information extraction and human computing for crowdsourced knowledge acquisition,” *2014 IEEE 30th International Conference on Data Engineering*, pp. 988–999, 2014. pages

[143] M. L. Jockers and G. Kirilloff, “Understanding gender and character agency in the 19th century novel,” 2016. pages

[144] T. Wu, J. Hu, and X. Zhu, “Character feature extraction for novels based on text analysis,” *2022 International Conference on Culture-Oriented Science and Technology (CoST)*, pp. 407–411, 2022. pages

[145] G. Xu, P. T. Isaza, M. Li, A. Oloko, B. Yao, A. Adebeyi, Y. Hou, N. Peng, and D. Wang, “Nece: Narrative event chain extraction toolkit,” *ArXiv*, vol. abs/2208.08063, 2022. pages

[146] D. Picca, “Semantic domains and supersense tagging for domain-specific ontology learning,” in *RIAO Conference*, 2007. pages

[147] N. Schneider, J. D. Hwang, V. Srikumar, J. Prange, A. Blodgett, S. R. Moeller, A. Stern, A. Bitan, and O. Abend, “Comprehensive supersense disambiguation of english prepositions and possessives,” *ArXiv*, vol. abs/1805.04905, 2018. pages

[148] M. Ciaramita and Y. Altun, “Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger,” in *Conference on Empirical Methods in Natural Language Processing*, 2006. pages

[149] A. S. Holloway and J. Ferguson, “Proceedings of the 14th annual conference of inebria,” *Addiction Science & Clinical Practice*, vol. 12, 2017. pages

[150] M.-A. Rizoiu, J. Velcin, and L. Eric, “Topic extraction for ontology learning,” 2011. pages

[151] A. Badawy, J. A. Fisteus, T. M. Mahmoud, and T. A. El-Hafeez, “Topic extraction and interactive knowledge graphs for learning resources,” *Sustainability*, 2021. pages

[152] M. L. Jockers and D. Mimno, “Significant themes in 19th-century literature,” *Poetics*, vol. 41, pp. 750–769, 2013. pages

[153] D. M. Blei, A. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2001. pages

[154] J. Laubrock and A. Dunst, “Computational approaches to comics analysis,” *Topics in cognitive science*, 2019. pages

[155] R. W. White, J. M. Jose, and I. Ruthven, “Comparing explicit and implicit feedback techniques for web retrieval: Trec-10 interactive track report,” in *Text Retrieval Conference*, 2002. pages

[156] G. Jawaheer, M. Szomszor, and P. Kostkova, “Comparison of implicit and explicit feedback from an online music recommendation service,” in *HetRec ’10*, 2010. pages

[157] O. Augereau, M. Iwata, and K. Kise, “A survey of comics research in computer science,” *J. Imaging*, vol. 4, p. 87, 2018. pages

[158] C. Guérin, C. Rigaud, A. Mercier, F. Ammar-Boudjelal, K. Bertet, A. Bouju, J.-C. Burie, G. Louis, J.-M. Ogier, and A. Revel, "ebdtheque: A representative database of comics," *2013 12th International Conference on Document Analysis and Recognition*, pp. 1145–1149, 2013. pages

[159] B. Klomberg, I. Hacımusaoğlu, and N. Cohn, "Running through the who, where, and when: A cross-cultural analysis of situational changes in comics," *Discourse Processes*, vol. 59, pp. 669 – 684, 2022. pages

[160] K. Aizawa, A. Fujimoto, A. Otsubo, T. Ogawa, Y. Matsui, K. Tsubota, and H. Ikuta, "Building a manga dataset "manga109" with annotations for multimedia applications," *IEEE MultiMedia*, vol. 27, pp. 8–18, 2020. pages

[161] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Conference on Empirical Methods in Natural Language Processing*, 2019. pages

[162] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015. pages

[163] G. van Rossum and F. L. Drake, "Python 3 reference manual," 2009. pages

[164] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019. pages

[165] J. Liu and C. Shah, "Interactive ir user study design, evaluation, and reporting," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2019. pages

[166] P. J. Ågerfalk and O. Eriksson, "Usability in social action: reinterpreting effectiveness efficiency and satisfaction," in *European Conference on Information Systems*, 2003. pages

[167] S. Parab and S. S. Bhalerao, "Choosing statistical test," *International Journal of Ayurveda Research*, vol. 1, pp. 187 – 191, 2010. pages

[168] V. Sundstedt, "Rabbit run: Gaze and voice based game interaction," 2009. pages

[169] T. Gossen, "Search engines for children - search user interfaces and information-seeking behaviour," 2016. pages

Declaration of Academic Integrity

I hereby declare that I have written the present work myself and did not use any sources or tools other than the ones indicated.

Datum: 29/06/2023 ....

(Signature)