TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

IT70LT
Friederike Mimi Freiin von Blomberg 221135IAFM

Evaluation of the Performance of a Visual Search Engine in Comparison to Text-based Search and Navigation

Master’s Thesis

Supervisor:	Daniel Beste
	MSc
Co-Supervisor:	René Pihlak
	MSc

Tallinn 2022

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

IT70LT
Friederike Mimi Freiin von Blomberg 221135IAFM

Visuaalse otsingumootori tulemuslikkuse hindamine võrreldes tekstipõhise otsingu ja navigeerimisega

Magistritöö

Juhendaja: Daniel Beste
MSc
Kaasjuhendaja: René Pihlak
MSc

Tallinn 2022

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, the literature and the work of others have been referenced. This thesis has not been presented for examination anywhere else.

Author: Friederike Mimi Freiin von Blomberg

2022-05-16

Abstract

An unplanned hour of downtime can cost large manufacturing companies up to 142 €/s. The fast availability of spare parts is therefore a key to efficient reactive maintenance. The goal of this thesis is testing the hypothesis that visual search is faster and more accurate than text search and navigation in spare part search. To that end, an overview of spare part management and its stakeholders is gained. The general functionality of search engines is described and differences between visual and text search are determined. Differences in task time and search success are to be retrieved through a quantitative usability test and post-test questionnaires. Unmoderated remote test sessions are designed and conducted on 54 experts and laypeople.

The obtained findings reveal that visual search is significantly faster than text search for both sample groups and more successful for the laypeople sample. The differences are tested for statistical significance. The confidence interval for the mean difference in task time ranges from 49.23 s to 161.21 s. A difference that users are likely to perceive. A statistically significant mean difference in success rate between visual and text search can only be determined for laypeople. The confidence interval for the mean difference in success rate ranges from 0.02 to 0.32. This raises the hypothesis that visual search can enable laypeople to find spare parts as accurately as experts by means of visual search. The results of the post-test questionnaires demonstrate that despite the uncertainties in statistical significance, the visual search experience is perceived as better by laypeople and experts. Therefore, the combination of statistical results and the qualitative user perception demonstrate the enormous potential of visual search in spare parts search.

This thesis is written in English and is 60 pages long, including 7 chapters, 22 figures, and 13 tables.

Keywords: visual search engine, text search, spare parts search, quantitative usability testing, efficiency, effectiveness

List of abbreviations and terms

3D	three-dimensional
API	application programming interface
ASA	American Statistical Association
CBIR	content-based image retrieval
ERP	enterprise resource planning
ID	identification
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
OCR	optical character recognition
OEM	original equipment manufacturer
QR	quick response
RFID	radio frequency identification
RIS	reverse image search
SERP	search engine results page
SKU	stockkeeping unit

Contents

List of Figures

List of Tables

1 Introduction

The proverb “time is money” governs maintenance management in the manufacturing industry. Recent surveys with large, multinational manufacturing companies reveal that an unplanned hour of downtime costs on average between 250 000 € and 511 000 € [1][2]. This can translate to up to 8600 € per minute and 142 € costs per second. In the automotive industry, the costs can exceed 1 250 000 € per hour [1]. These costs include among others the costs of wages and salaries involved for reactive maintenance and the expediting of emergency spare parts [1]. According to Biedermann, working time studies show that maintenance staff spend an average of 10 % of their daily working time identifying suitable spare parts and finding them in the warehouse [3]. Considering that labour cost amounts to up to 43 % of direct maintenance costs [4], a significant amount of maintenance costs is being spent on spare part search. Shortcomings in the identification and search process of spare parts affect stakeholder along the whole supply chain of spare parts [5].

This thesis evaluates the effect of a new dimension to spare part search: visual search. To this end, the decisive metrics for the usability of search engines are examined and a quantitative usability test is conducted. For this purpose, the expertise of maintenance technicians and spare parts managers from the manufacturing industry is needed. The goal of the thesis is to test the hypothesis that visual search is faster and more accurate than text-based search and navigation. Thereby, a new benchmark for further research is targeted. The framework of this thesis is summarised in Figure 1.

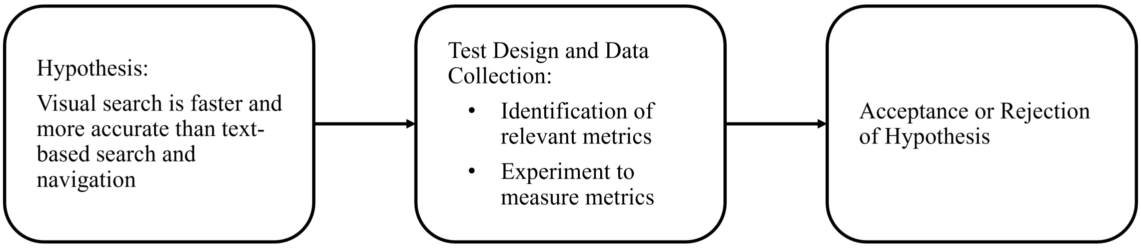

Hypothesis:
Visual search is faster and more accurate than text-based search and navigation
Test Design and Data Collection:
• Identification of relevant metrics
• Experiment to measure metrics
Acceptance or Rejection of Hypothesis

Figure 1. Framework of Thesis.

The thesis is structured along the following chapters:

- First, in Chapter 2, a brief introduction of spare parts management systems is given and a basic understanding of the technical functionality and the use cases of search engines is created.

- In Chapter 3 the state of the art metrics to evaluate search engines from the user experience perspective are presented.
- Chapter 4 covers the methodology of quantitative usability testing that is applied in this thesis. First, the selection of the search query data, the test search engines and the test users is defined. Secondly, the method for the analysis of the resulting search metrics and the approach to statistical testing is explained.
- In Chapter 5 the empirical results of the benchmark test between visual and text-based search are presented and statistically analysed. In addition, the collected qualitative feedback data from the user testing is summarised.
- In Chapter 6 the applied method and the obtained results are critically reflected. Subsequently, conclusions are drawn and further need for research is discussed.
- Chapter 7 provides a summary of the thesis.

2 Background

This chapter provides an introduction into spare parts management and introduces the concept of search engines. Thereby, the scope of examination of the present thesis is defined as illustrated in Figure 2. Furthermore, terms for better understanding of the thesis are introduced and specified. In the first section, an overview of spare parts management is given. A summary of spare parts management systems deployed for spare part search is presented and its stakeholders are defined. Subsequently, an introduction into search engines is given and their functionality is briefly explained. Thereby, differences between text and visual search are outlined.

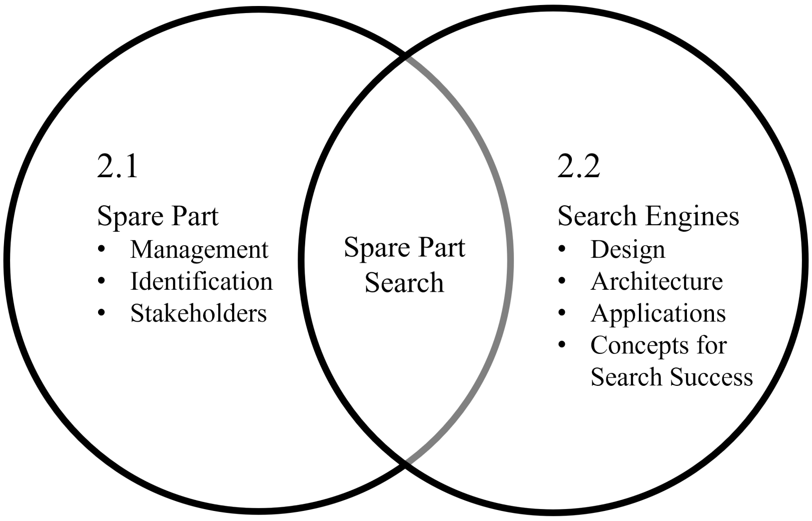


Figure 2. Structure of chapter 2 'Background'.

2.1 Spare Parts Management

According to Teixeira et al., spare parts management is a function of maintenance management that aims to support maintenance activities, giving real-time information on the available quantities of each spare part and adopting the inventory policies that ensure their availability when required, minimizing costs [6]. The components of a spare part system and their relationships are illustrated in Figure 3 according to Biedermann. The spare part inventory is primarily shaped by the system and the technology that manufacturing companies deploy. Secondarily, the customer or machinery, determines the demand of spare parts through maintenance activities. The inventory policies identify which spare parts are ordered from the supplier.[3]

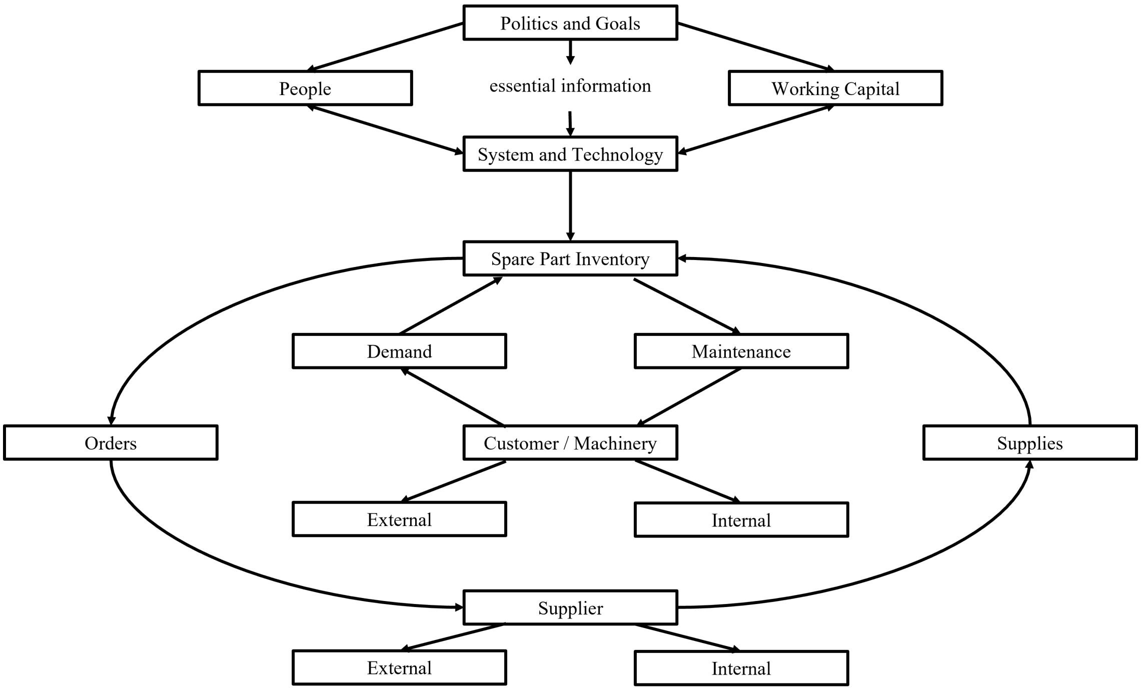


Figure 3. Components of spare part system [3].

Reactive maintenance and therefore **efficient spare parts management is crucial** since large manufacturing companies have unplanned downtime costs of 250 000 € to 511 000 € **per hour** [2][1]. In the automotive industry, the costs are even larger at 1 250 000 € per hour. Although these figures span a large range, they indicate the significant magnitude of cost due to unplanned downtime.

2.1.1 Spare Parts Management Systems

In most manufacturing and production systems, an enterprise resource planning (ERP) system is used to manage and monitor spare parts in stock as well as in- and outflows. In all ERP systems, the stock database contains basic data on each item, including a unique item number, an item description, the current quantity in stock, the ordered quantity and the minimum stock level. **Today's ERP systems accommodate text-based search with different levels of complexity.** According to Biedermann, it is important that in the event of plant malfunctions, maintenance can identify the required spare parts with a high degree of accuracy and schedule them from the warehouse. It is known from working time studies that maintenance staff spend an average of 10 % of their daily working time identifying suitable spare parts and finding them in the warehouse.[3]

2.1.2 Spare Part Identification Systems

According to Biedermann, identification systems are becoming increasingly important in spare parts management. They are deployed in the spare parts management systems concerning the stock, the transfer or retrieval, the transport, and also the repair of spare parts. Identification systems support both the organisational processes and the allocation of spare parts to orders.[3] The most common technologies for spare part identification are serial numbers on the spare part itself, bar codes, 2D data codes and radio frequency identification (RFID) labels. Bar codes and 2D data codes such as the quick response (QR) code are easily applicable as stick-on labels.[7] They can easily be read by laser-based scanners or any smartphone device with the appropriate application. However, many spare parts do not possess any identifier at all. Therefore, traditional paper documentation and catalogues need to be searched through in order to identify them. An additional challenge attached to bar code and QR code labels is their sensitivity to dirt and destruction in an industrial environment, leaving them unreadable. RFID labels, on the other hand, can be read without line-of-sight access. However, the decisive disadvantages of the aforementioned technologies concerning spare part identification are that they cannot be applied to small spare parts and suffer through wear and tear.[3]

2.1.3 Stakeholder of Spare Part Search

According to Oumaima et al., the supply chain of spare parts is the intersection between the supply chain, after-sales and maintenance services [8]. Spare parts search is conducted on both sides of the spare part supply chain, i.e. on the supplier side as well as on the consumer side. On the consumer, or spare part user side, spare part search may be allocated as a cross-divisional supply chain function. The stakeholders that undertake spare part search regularly may come from different divisions, i.e. procurement or manufacturing.[3] Typical stakeholders on the spare part consumer side are maintenance and service technicians, as well as maintenance engineers. On the supplier side, spare part search is conducted mainly by the customer and technical support. However, stakeholders on both sides may also come from cross-divisional spare part logistic and management functions.[3] The decentralized management configuration of spare parts causes the lack of information sharing between technicians, which can generate significant costs and low quality of service [8].

Besides the direct stakeholders, spare parts search affects multiple indirect stakeholders, since downtime of machinery can cause a chain of delays. These include customers of spare part consumers who are waiting for their products or original equipment manufacturer (OEM) sales since machinery sales might depend on the quality of spare parts service.

2.2 Search Engines and Information Retrieval

Information retrieval includes structuring, analyzing, organizing, storing, searching and retrieving information [9]. The focus of this thesis lies on the search and retrieval of spare part information. Thus, in the following, the background information for a general understanding of search engines is given. The first section outlines the fundamental concepts of search engine design. In the second section, the generic architecture of search engines and the differences between text and visual search are presented. The third section addresses types of applications of search engines that are relevant to the understanding of this thesis. The forth section provides a brief overview of the role of navigation and the importance of search engine results pages (SERPs) in search success.

2.2.1 Search Engine Design

Search engine design is based on the fundamental themes of information retrieval. According to Croft et al., these themes are distributed among the following core topics:

- Relevance
- Evaluation
- Information need

The concept of **relevance** defines whether a search result contains the information a user is looking for when submitting a search query to a search engine. Relevance can be categorised into two types: *topical relevance* and *user relevance*. Topical relevance signifies that the search result file is on the same topic as the query. User relevance, however, takes the subjective perception into account, i.e. whether a user perceives the result as relevant or not. Within the scope of this thesis, the focus lies on user relevance. The second key theme in information retrieval is **evaluation**. Evaluation concerns the development and deployment of metrics to measure the quality of the retrieved information. The evaluation of search engines is the key theme of the present thesis and is focused on in Chapter 3.

The third core topic of information retrieval is the **information need**. An information need is the underlying cause of the query that a person submits to a search engine. The information need is presented to the search engine in the form of a query. Especially text queries are often poor descriptions of what the user is actually searching for. According to Croft et al., numerous studies are concentrating on the development of techniques to help people express their information needs. This emphasises the poor usability of text search in terms of expressing the information need.[9]

2.2.2 Search Engine Architecture

According to Croft et al., a search engine is a software system that applies information retrieval techniques to large-scale databases, i.e. a collection of either structured or unstructured data [9]. In other words, a search engine is a computer program to find answers to queries in a collection of information, which might be a library catalog or a database [10]. The two primary objectives of a search engine are:

- effectiveness (retrieve the most relevant set of results for a search query)
- and efficiency (retrieve the results for a search query as quickly as possible).

These determine the architecture of a search engine and will be presented in further detail in chapter 3. The architecture of search engines consists of software components, their interfaces and the relationships between the components. An outline of the main components and their tasks is presented in the following section.

2.2.3 Indexing and Query Process

The software components fulfill two main tasks, the *indexing process* and the *query process*, i.e. getting data into the search engine and out again during search. The indexing process creates the data structure that the query process accesses during the search process. Thereby, the query process matches a user's query to the results of the indexing process to produce a ranked list of files.[9]

The **indexing process** consists of three main parts, i.e. *data acquisition*, *data transformation*, and *index creation* as depicted in Figure 4. The data acquisition component identifies data files that will be searched and stores it in a data store, i.e. the data collection. This

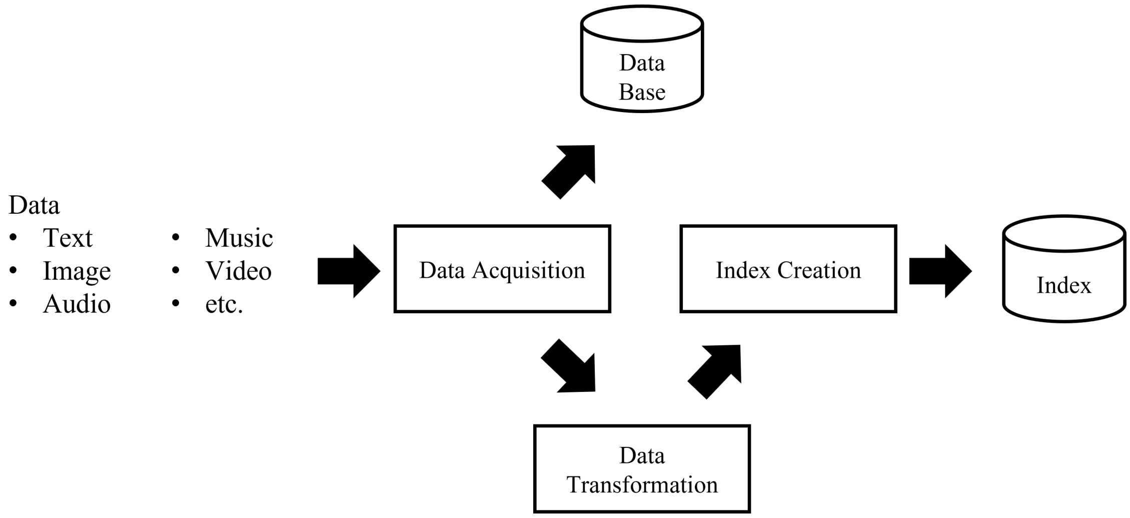


Figure 4. Indexing Process according to Croft et al.[9].

data store contains the data files and also *metadata* for all data. Metadata is information about data files that is not part of the data file content. Metadata can contain a multitude of information, such as the type (e.g. text, image or web page), the structure, or the size of a file. Data files stored through the data acquisition process are then transformed by the data transformation component into *features*, which are called *index terms* for text files. Index terms, or features, are the representatives of the data file content. The index creation component takes the index terms and creates data structures to enable fast searching. The main objective in index creation is efficiency since indexes need to be updated as fast as possible when new data is added to the data store.[9][11]

The **query process** also consists of three main components, i.e. *user interaction*, *ranking* and *evaluation* and is illustrated in Figure 5. The user interaction components provides the interface to the user. Upon query entry, the user interaction component transforms it into features. The user interaction component also deploys multiple techniques to refine the query so that it better represents the information need of the user. This component also accepts the ranked list of returned data files from the search engine and organises the list into a search engine result page. To this end, it resorts to the data base to get information from the respective data file. The heart of a search engine is represented by the ranking component. The ranking component or retrieval model compares or matches the feature vectors to those in the database using a previously defined similarity (also distance) measure. There is a match if the distance between the query vector and a vector representing an indexed image is less than or equal to a threshold. The efficiency of the ranking correlates to the indexes and the effectiveness to the retrieval model. Finally, the evaluation

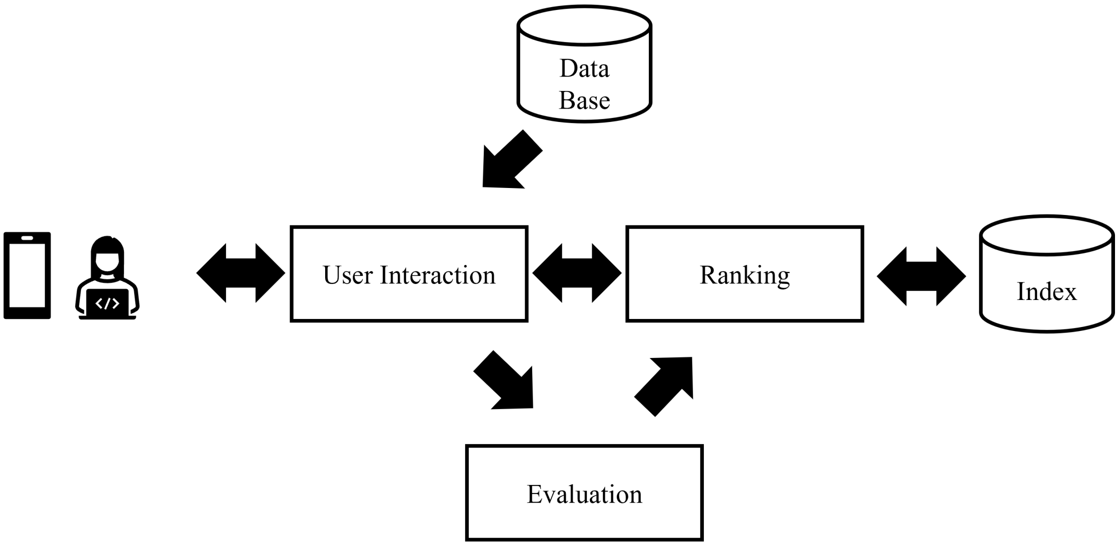


Figure 5. Query Process according to Croft et al.[9].

component provides measurements and monitoring of efficiency and effectiveness.[9]

Despite the obvious **differences between text and visual search** that are outlined in the following, the basic algorithms and functions are very similar and their architectures are based on the indexing and query processes depicted in Figure 4 and Figure 5. For illustration purposes, the complete visual search process is depicted in Figure 6. Text and visual search differ primarily by the type of query data and thus also the data transformation process. In text search the query data consists of text files, while visual search or reverse image search takes images as input and returns results related to the query image [12]. Visual search is a type of content-based image retrieval (CBIR)[13], which means that it has to solve the intrinsic problem of describing an image mathematically in order to create a searchable index [14]. According to Datta et al., the abstract mathematical description of an image, for retrieval purposes, is referred to as its signature. The construction of image signatures corresponds to the extraction of features, or the extraction of index terms for text files. Features in terms of visual search are defined to be visual properties of an image, either globally for the entire image or locally for a small group of pixels. The most commonly used features are color, texture, shape, and salient points in an image.[14] Similar to the text-based search, the result of the feature vector matching process is a ranked list of images.[13]

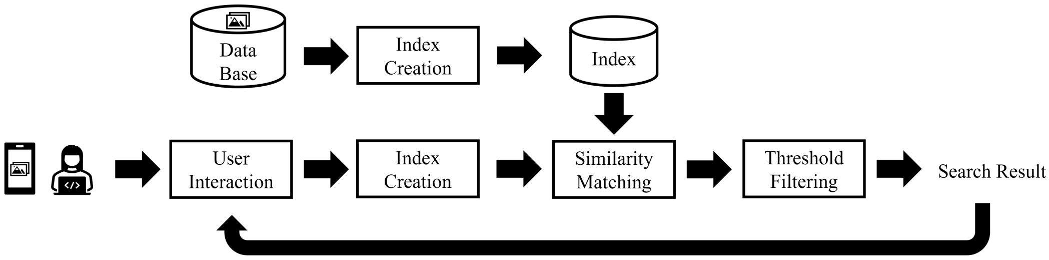


Figure 6. Reverse image search (RIS) based on Gaillard et al.[13].

2.2.4 Applications of Search Engines

Search engines are deployed over a range of tasks and applications. The most common application of information retrieval is the search on the internet, also known as *Web Search*. Another well-known search engine is *desktop search*, where the information corpus are the files stored on an individual computer [9]. There are several other applications of search engines. Within the scope of this thesis, the focus lies on enterprise and site search, which are deployed for spare part search. *Enterprise Search* is defined as the information retrieval among a large variety of both structured and unstructured computer files across a distributed corporate intranet. The corporate intranet may consist of various different sources such as web pages, reports, spreadsheets and structured data in corporate databases.[9][15] Enterprise search differs a lot from searching the web since when searching an intranet, commonly a “right” result exists as a specific file. In contrast, when searching the web, it is the best matching set of web pages that are relevant. According to Mukherjee et al., enterprise search may be more difficult, since **finding the right answer is often more difficult than finding the best answer**.[16] *Site Search*, often also e-commerce search, is similar to Web Search but is restricted to the web pages at a given website. Site search enables companies to implement search engines on their websites. The providers of site search only index the respective website through an application programming interface (API) and implement a search box on the website [17].

2.2.5 Concepts Contributing to Search Success

Many enterprise and site search engines combine their text search interface with a faceted **navigation** structure since it **allows a user to elaborate the search query progressively**, learning from the available options [18]. Most of the data collections are semi-structured. Faceted search is the combination of faceted navigation of structured content and text search applied to unstructured text content. User studies demonstrate that faceted search provides more effective information-seeking support to users than conventional search. Faceted search has become increasingly prevalent in online information access systems, particularly for e-commerce and site search.[18]

As outlined in Section 2.2.1, the query process ends on the ranked list of results, also called SERP. According to a study by Oulasvirta et al., the presentation of search results in general affects the choice and satisfaction of users significantly. The study proved that an increasing recall, i.e. the proportion of relevant documents that are retrieved, can actually work counter to user satisfaction if it implies choice from a more extensive set of result items.[19][9] Regarding reverse image search, according to Datta et al., the presentation of search results is one of the most important factors in the acceptance and popularity of an image retrieval system [14].

3 State of the Art

This chapter sets out the concept of search engine usability evaluation. Hereby, the scope of parameters to be examined in the quantitative usability testing of visual against text search is defined. Terms for better understanding of the thesis are introduced and specified. In the first section search usability evaluation metrics are examined and a selection is made for further deployment. In the second section a brief overview of existing evaluations for text-based search and visual search is presented and benchmark deficits are elaborated.

3.1 Search Engine Usability Evaluation

According to Croft et al., evaluation is essential to understanding whether a search engine is being used effectively in a specific application. Therefore, evaluation is key in enhancing the performance of search engines.[9] In the present thesis the focus lies on the **usability dimensions** of search engine performance. The International Organization for Standardization (ISO) 9241.11 standard defines usability as “the extent to which a product can be used by specified users to achieve specified goals with **effectiveness**, **efficiency** and **satisfaction** in a specified context of use”.[20] These concepts are introduced in the following sections. The respective usability metrics are defined in the ISO/International Electrotechnical Commission (IEC) 25 022:2016 standard *Systems and software engineering - Systems and software quality requirements and evaluation (SQuaRE) - Measurement of quality in use* and will be elaborated on below [21].

3.1.1 Usability Metrics for Effectiveness

According to ISO/IEC 25 022:2016, effectiveness is the accuracy and completeness with which users achieve specified goals [21]. According to Sauro et al., the key effectiveness metrics are usability problems, errors and completion rates. Usability problems have names, a description, and often a severity rating that takes into account the observed problem frequency, i.e. the number of errors, and its impact on the user. Errors are any unintended action, slip, mistake, or omission a user makes while attempting a task. Error counts can go from 0 (no errors) to technically infinity. Completion rates, also called **success rates**, are the most fundamental of usability metrics. They are typically collected as a binary measure of task success (coded as a 1) or task failure (coded as 0). Within the scope of the present thesis, the focus lies on the **success rate of a search task** as a metric for effectiveness of search engines.[22]

3.1.2 Usability Metrics for Efficiency

According to ISO/IEC 25 022:2016, efficiency encompasses the resources expended in relation to the accuracy and completeness with which users achieve goals [21]. According to Sauro et al., efficiency is measured in terms of **task time**, i.e. the amount of time that a user spends on an activity. Most often the task time is the amount of time that it takes a user to successfully complete a predefined task scenario but it can also be total time on a web page or call length. Task time, or duration, is typically reported as an average.[22] According to Sauro et al., the task duration can be measured and analyzed in the following ways [22]:

- Task completion time: Time of users who completed the task successfully.
- Time till failure: Time on task until users give up or complete the task incorrectly.
- Total time on task: The total duration of time users are spending on a task.

Within the scope of this thesis, task time refers to the **total time on a search task**. Efficiency is a way of combining task success and task times into a single measure that represents task success per unit of time [23]. However, within the scope of this thesis, the metrics are analysed separately from one another.

3.1.3 Usability Metrics for Satisfaction

According to ISO/IEC 25 022:2016, satisfaction signifies the comfort and acceptability of use of a software, e.g. search engine [21]. Satisfaction can be measured by means of questionnaires. According to Sauro et al., questionnaires that measure the perception of the ease of use of a system can be completed immediately after a task (post-task questionnaires), at the end of a usability session (post-test questionnaires), or outside of a usability test [22]. Within the scope of the present thesis, **post-task questionnaires** are deployed.

3.2 Benchmarking Deficits

Numerous studies have focused on search engine (i.e. both text-based and RIS) evaluation based on the ISO/IEC 25 010:2011 standard *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*, for example: [24], [25] and [9]. This standard encompasses usability among the main software quality attributes (i.e. functionality, reliability, performance, maintainability, potability and reusability). According to this standard, the evaluation of usability metrics concentrates on search engine architecture metrics such as accuracy, precision or response time and are typically done in tightly defined experimental settings [9].[26]

However, according to Croft et al., search is an interactive process involving different types of users with different information problems. In this environment, the performance of search engines will be affected by many factors, such as the interface used to display search results and query refinement techniques, such as query suggestion and relevance feedback.[9] At present, there is a research deficit for quantitative comparative evaluations of the usability of search engine performance regarding different types of search engines. The present thesis presents an attempt at this holistic evaluation of the performance of visual search against text-based search and navigation.

4 Methodology

In this chapter, the methodology to compare the performance of visual search to text search and faceted navigation is set. The aim is to reject statistically the null hypothesis that there is no difference in performance between visual and text search. Hereby, a **comparative empirical** approach to usability analysis is chosen. A summary of the applied method is given below and illustrated in Figure 7.

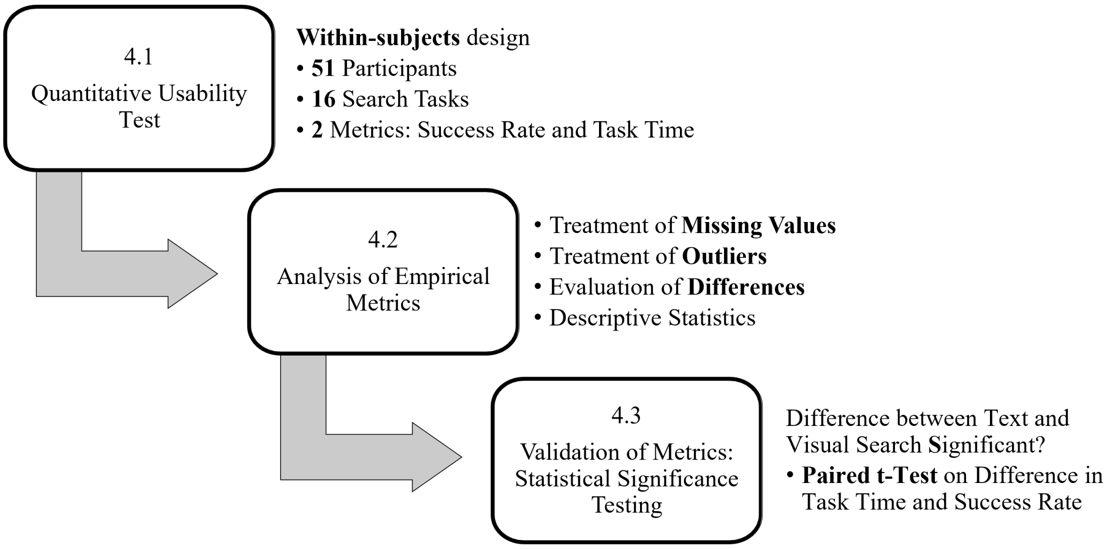

4.1
Quantitative Usability Test
Within-subjects design
• 51 Participants
• 16 Search Tasks
• 2 Metrics: Success Rate and Task Time
4.2
Analysis of Empirical Metrics
• Treatment of Missing Values
• Treatment of Outliers
• Evaluation of Differences
• Descriptive Statistics
4.3
Validation of Metrics: Statistical Significance Testing
Difference between Text and Visual Search Significant?
• Paired t-Test on Difference in Task Time and Success Rate

Figure 7. Applied method for evaluating the performance difference between visual and text-based search.

In the first section, a method for data collection, quantitative usability testing, is chosen and introduced. Hereby, the **within-subjects test design** is defined and its application on search engine comparison is elaborated on. The subsequent definition of the evaluation corpus and the selection of participants for the sample sets out the scope of examination. Two search query test sets, 16 spare parts each, are defined. Two sample groups are selected: 1) maintenance technicians and spare parts managers and 2) laypeople. Concluding the first section the deployed visual and text-based search engines are examined.

The second section addresses the approach to pre-processing and applying descriptive statistics to the metrics obtained in the quantitative usability test. The calculation of **difference metrics** and their examination regarding central tendencies and measures of variability is outlined. This examination is conducted for both sample groups.

The third section covers the method to test whether the difference in metrics between the search engines is greater than what would be expected from chance. By determining the statistical significance, the difference that may likely exist in the untested population may be determined.

4.1 Quantitative Usability Testing

The goal of the quantitative or summative approach to usability testing is to describe the usability of an application using metrics [22]. Within the scope of the present thesis, the metrics introduced in chapter 3.1 are examined: **success rate, task time and satisfaction**. According to Sauro et al., collecting multiple metrics in a usability test is advantageous because this provides a better picture of the overall user experience than any single measure can [22].

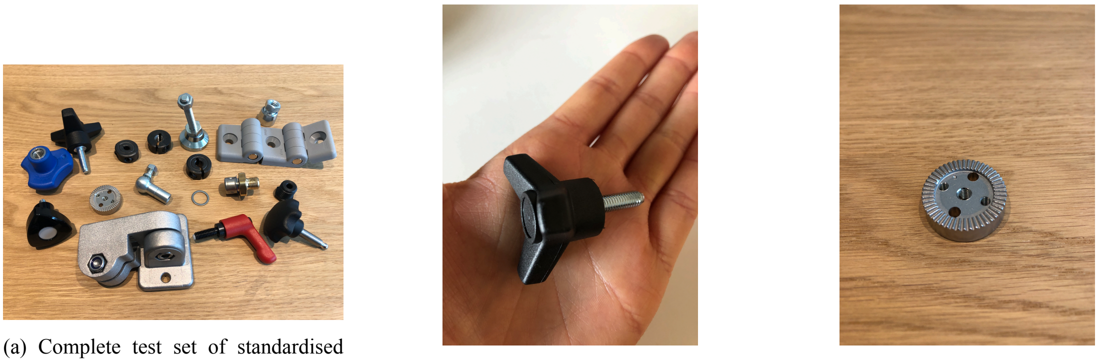
4.1.1 Evaluation Corpus

The evaluation corpus, i.e. the data set deployed for the test, is based on the products and services catalogue from a standard spare part manufacturer. The data set consists of 34 534 standard spare parts with a total number of 34 817 images. The spare parts are distributed across the following main product categories: *Operating Parts*, *Clamping Parts*, *Machine Parts*. The evaluation corpus comprises the following data to each spare part: full name of spare part, product identification (ID) (also stockkeeping unit (SKU)), attributes of the part, a dimensions image, the article group title, group ID, a photo of the part, the **assignment to five level categories** and a three-dimensional (3D) model of the part.

According to Croft et al., the evaluation corpus in information retrieval is unique in that the queries and relevance judgments for a particular search task are gathered in addition to the documents.[9] In the case of the present experimental setup the relevance judgements for each query are conducted by the same three people. The scope of relevance judgements is very small since there are only one or two files, i.e. spare parts, relevant to each query.

Query Test Set

16 spare parts are randomly selected from the product catalogue across different categories and ordered. The complete set of spare parts that represent the test set is depicted in Figure 8a. Two query images are shown in Figure 8c and 8b for illustration purposes.



(a) Complete test set of standardised spare parts.

(b) Example of query image: part no. 8.

(c) Example of query image: part no. 2.

Figure 8. Set of spare parts deployed for quantitative usability test.

The selection of various spare parts that constitutes the two query test sets A and B is listed in Table 1. The spare parts for the query set are selected randomly across nine different first-level categories out of a total of 17 first-level categories.

Table 1. List of all spare parts of test query sets.

No.	Full name of spare part
1	GN 706.2 Semi-Split Shaft Collars
2	GN 187.4 Stainless Steel Serrated Locking Plates
3	GN 159.1 Double Hinges
4	GN 5339 Triangular Knobs with Threaded Stud
5	GN 5342 Three-Lobed Knobs
6	GN 346 Ball Joint Thrust Pads
7	GN 884 Breather Filters
8	GN 5330 Three-Lobed Knob Screws
9	GN 306 Adjustable Hand Levers
10	GN 343.2 Leveling Feet
11	DIN 71802 Angled Ball Joints
12	GN 707.2 Split Shaft Collars

Continues...

Table 1 – *Continues...*

No.	Full name of spare part
13	GN 706.3 Threaded Shaft Collars
14	GN 6336.11 Star Knobs
15	GN 412.2 Positioning Bushings
16	GN 281 Swivel Clamp Connector Joints

The nine first-level categories that the spare parts for the test set are picked from read as follows: 1. Mounting, positioning, levelling with screws, clamping and supporting elements, 2. Hinging, latching, locking of doors and covers, 3. Tensioning, clamping with knobs, 4. Controlling, venting, sealing of liquids and gases, 5. Tensioning, clamping, switching with levers, 6. Installing, lifting, dampening with leveling feet, lifting gear and rubber elements, 7. Moving, transferring, connecting with shafts and joints, 8. Indexing, locking, blocking with pins and ball-shaped elements, 9. Connecting, assembling with clamping and connecting elements.

4.1.2 Sampling of Test Participants

For both sample groups, a sample size of 30 or greater is targeted since according to the Central Limit Theorem, sample means tend to have normal distributions at this sample size [27]. The two governing concepts of sample selection are **representativeness** and **randomnessSauro**. According to Sauro et al., the most important thing in user research is that the sample of users that metrics are obtained from *represents the population* about which statements are intended to be made. Randomness is desirable since statistical tests in general are based on the assumption that users are sampled randomly from a population. However, according to Sauro et al., it is more important to have a less-than-perfectly random sample from the right population than if you have a perfectly random sample from the wrong population.[22]

Sampling Technique

The participants are recruited by means of two non-probability sampling techniques: **convenience sampling** and **snowball sampling**. Convenience sampling refers to sampling from a group of people that are easy to reach. Snowball sampling is a type of convenience sampling in which those participants invited invite other participants and so on to create a pyramid effect.[23] Companies from different manufacturing industries are contacted to recruit participants. To increase response rates, participants are offered an incentive. The qualification of participants is examined on the basis of a registration process and a **screener survey**. The questions of the screener survey are listed in Table 2.

Table 2. Screener survey for potential test candidates.

Question Nr	**Screener Question**
Question 1	What is your profession?
Question 2	Which industry do you work in?
Question 3	How much time per week do you spend identifying spare parts?
Question 4	Please indicate your age range. ■ Under 30 years ■ 30 - 40 years ■ 40 - 50 years ■ Above 50 years
Question 5	Do you agree that a recording of your screen will be made during the study participation? This serves as a random check for the test implementation. The recordings are automatically deleted after one month. ■ YES ■ NO

Composition of Sample Groups

As examined in chapter 2.1.3, spare part search is conducted mainly by maintenance and service technicians, spare parts managers and customer support staff. Based on this population, the panel of participants for two sample groups is recruited. The **first sample group consists of** 23 **maintenance technicians and spare parts managers** and in the following will be referred to as the *expert sample*. The **second sample group consisting of** 30 **mechanical engineers and laypeople** represents the customer support staff and will be referred to as *support sample*.

In the following a brief overview of the composition of the two sample groups is provided. The *support sample* group consists of 50 % mechanical engineers and 50 % laypeople. The *expert sample group* is employed across the following industries: cable manufacturing, testing equipment manufacturing, automation technology, general industry maintenance, rail traffic, wallpaper manufacturing, tooling manufacturing, fuel cell manufacturing and metal processing. The distribution across the industries is illustrated in Figure 9. Figure 10 depicts the composition of both sample groups by age. The average amount of time the members of the expert sample group spend on spare part identification each week is illustrated in Figure 11.

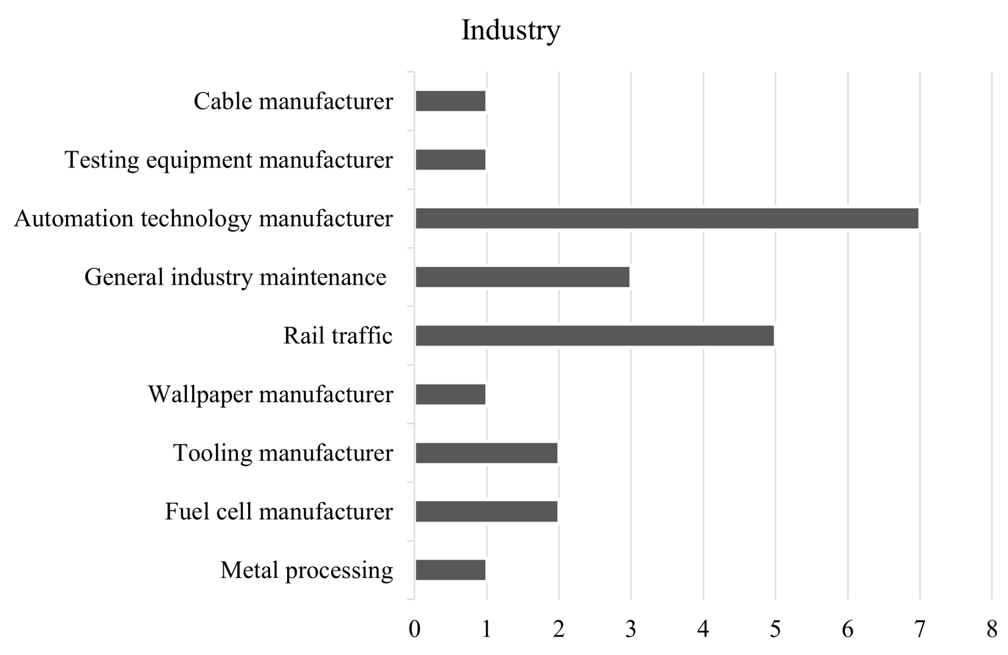


Figure 9. Distribution of industries from which spare parts manager sample drawn.

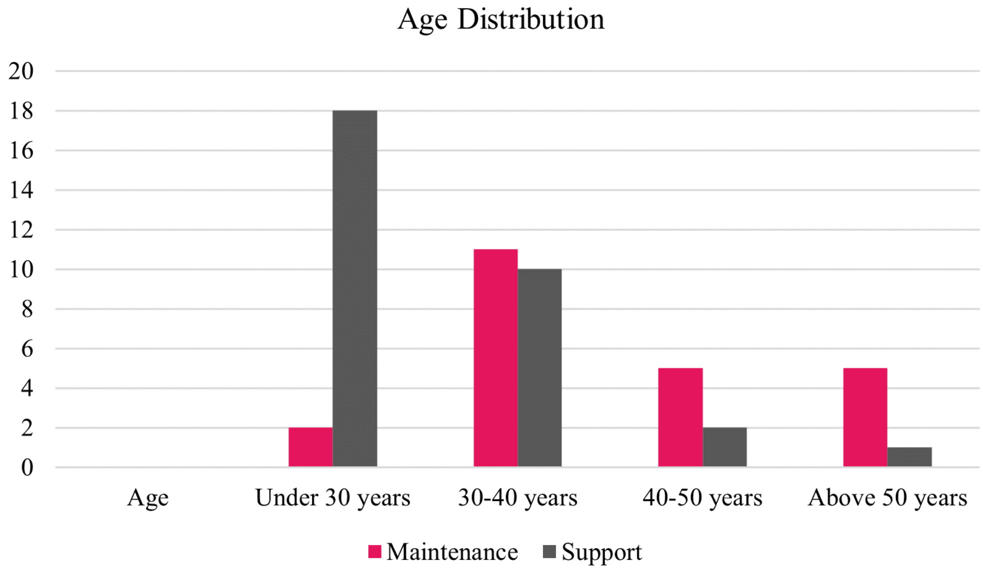


Figure 10. Age composition of both sample groups.

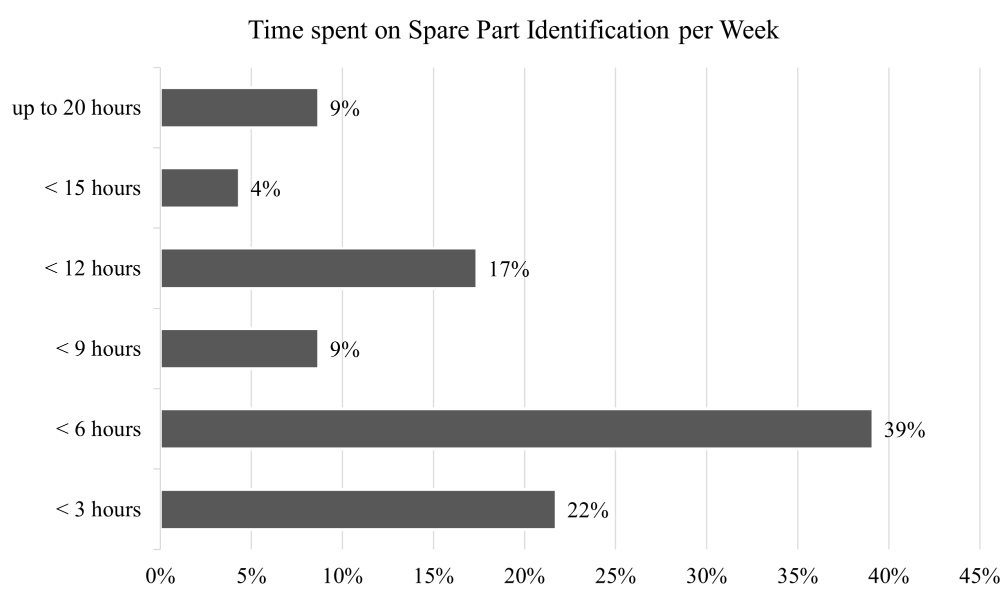


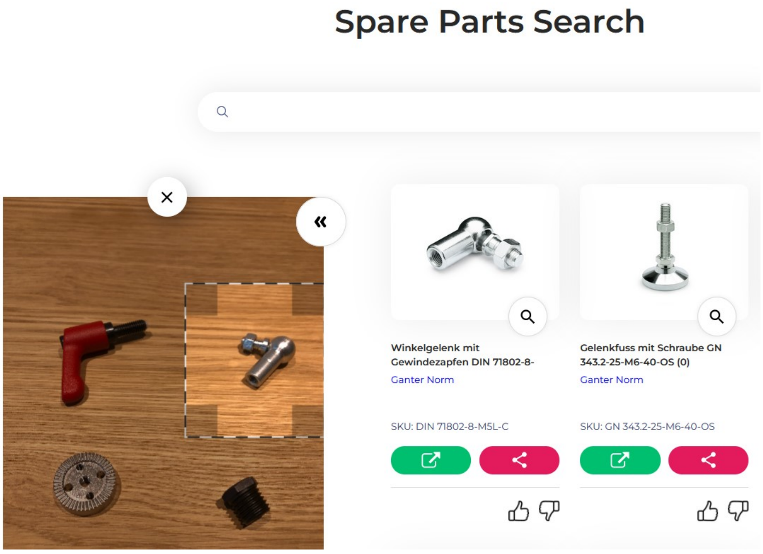
Figure 11. Time spent on spare part identification per week by spare parts manager sample.

4.1.3 Visual Search Engine

For the search test experiment, a visual search engine from *nyris*, a software development company from Berlin (Germany), is applied in the test. The visual search engine supports the following main functionalities:

- generic object detection
- image similarity search
- text search

Object detection is provided by the Google Cloud Vision API. This generic object detection allows the nyris similarity search to compute the image similiarity between localised objects in the query and the database instead of using the whole image. Optical character recognition (OCR), however, is not implemented in the search system. During the image similarity search process, 384-dimensional feature vectors of the images are created by means of an unsupervised neural network. Unsupervised refers to the concept that the training data is not classified [11]. The neural network is based on the DINO infrastructure by MetaAI. The visual search is supported by a text search, i.e. the visual search results can be refined through keyword search. The text search is provided by the Algolia API. Since in fact, two search engines are applied in the test system, the data has to be imported and indexed in both environments.


Spare Parts Search

Winkelgelenk mit Gewindezapfen DIN 71802-8-
Ganter Norm
SKU: DIN 71802-8-M5L-C

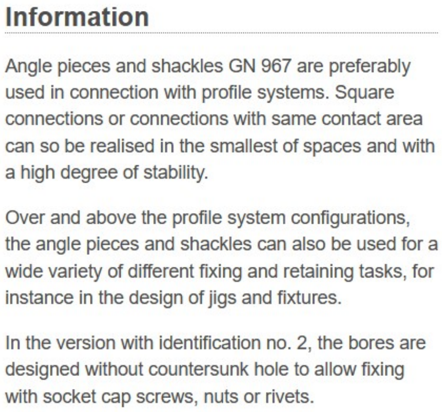
Gelenkfuss mit Schraube GN 343.2-25-M6-40-OS (0)
Ganter Norm
SKU: GN 343.2-25-M6-40-OS

Figure 12. Interface of nyris spare part search engine [28].

The search engine is optimised through upload of two to three additional photos per SKU into the index (for the indexing process see chapter 2.2.5). The neural network deployed is **not** trained on these photos. Additionally, the search is optimised through synthetic data, i.e. the rendered CAD data is added to the index. For convenience, the search engine deployed for comparison will be called *visual search* in the following.

4.1.4 Benchmark Search Engine

The search engine deployed for comparison is the **text-based e-commerce search** on the website of the standard spare part manufacturer. Thereby, the identical data basis is ensured for both search systems, i.e. the product catalogue of the spare part manufacturer. However, different to the visual search system, the on-site text-based search also **indexes the product description** of the spare part. An example of a product description is illustrated in Figure 13.



Information

Angle pieces and shackles GN 967 are preferably used in connection with profile systems. Square connections or connections with same contact area can so be realised in the smallest of spaces and with a high degree of stability.

Over and above the profile system configurations, the angle pieces and shackles can also be used for a wide variety of different fixing and retaining tasks, for instance in the design of jigs and fixtures.

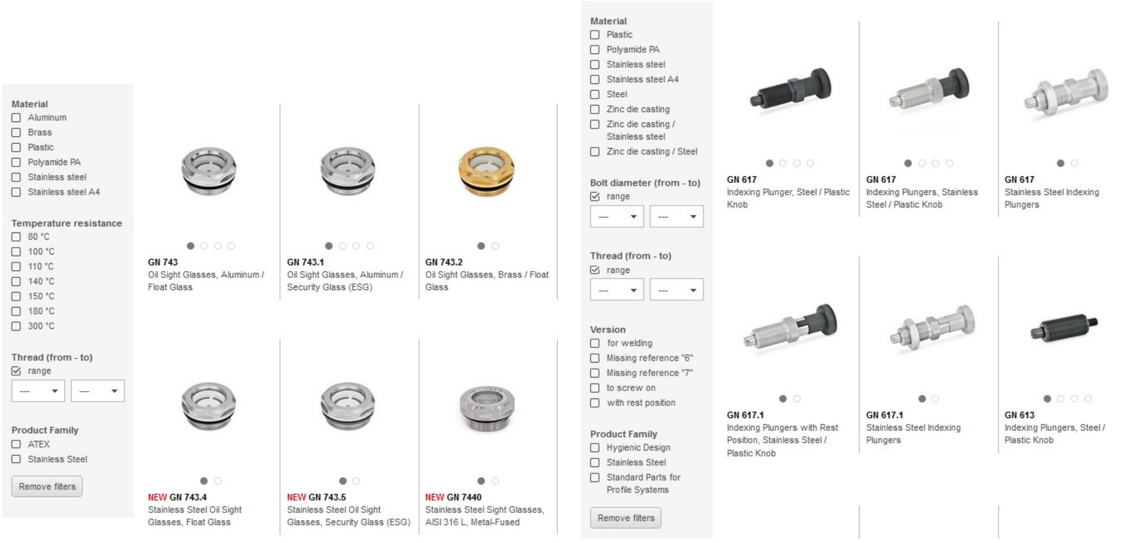
In the version with identification no. 2, the bores are designed without countersunk hole to allow fixing with socket cap screws, nuts or rivets.

Figure 13. Example of product description for spare part *angle pieces* [29].

The product catalogue search functionalities encompass **multiple keyword search**, **number range search**, **word variations (plurals, etc.)**, **filtering** and **faceted navigation**. Filtering options and the number range search are illustrated by the examples given in Figure 14a and Figure 14b. Along the text-based search functionalities, the product catalogue search supports **faceted navigation**. As outlined in chapter 2.2.5, faceted navigation allows the user to elaborate a query progressively seeing the effect of each choice in one facet on the available choices in other facets. The faceted navigation architecture is implemented along five category levels, three of those are visible to the user. The distribution of categories across the three main product categories reads as follows:

1. Operating Parts: four first-level categories, 29 second-level categories
2. Clamping Parts: four first-level categories, 24 second-level categories
3. Machine Parts: nine first-level categories, 71 second-level categories

Natural language processing techniques that involve syntactic or semantic analysis are not sup-



Material
- □ Aluminum
- □ Brass
- □ Plastic
- □ Polyamide PA
- □ Stainless steel
- □ Stainless steel A4

Temperature resistance
- □ 80 °C
- □ 100 °C
- □ 110 °C
- □ 140 °C
- □ 150 °C
- □ 180 °C
- □ 300 °C

Thread (from - to)
- ☑ range

--- ▾ --- ▾

Product Family
- □ ATEX
- □ Stainless Steel

Remove filters

GN 743
Oil Sight Glasses, Aluminum / Float Glass

GN 743.1
Oil Sight Glasses, Aluminum / Security Glass (ESG)

GN 743.2
Oil Sight Glasses, Brass / Float Glass

NEW **GN 743.4**
Stainless Steel Oil Sight Glasses, Float Glass

NEW **GN 743.5**
Stainless Steel Oil Sight Glasses, Security Glass (ESG)

NEW **GN 7440**
Stainless Steel Sight Glasses, AISI 316 L, Metal-Fused

(a) Illustration of filtering on *oil sight glasses* from website of spare part manufacturer [29].

Material
- □ Plastic
- □ Polyamide PA
- □ Stainless steel
- □ Stainless steel A4
- □ Steel
- □ Zinc die casting
- □ Zinc die casting / Stainless steel
- □ Zinc die casting / Steel

Bolt diameter (from - to)
- ☑ range

--- ▾ --- ▾

Thread (from - to)
- ☑ range

--- ▾ --- ▾

Version
- □ for welding
- □ Missing reference "6"
- □ Missing reference "7"
- □ to screw on
- □ with rest position

Product Family
- □ Hygienic Design
- □ Stainless Steel
- □ Standard Parts for Profile Systems

Remove filters

GN 617
Indexing Plunger, Steel / Plastic Knob

GN 617
Indexing Plungers, Stainless Steel / Plastic Knob

GN 617
Stainless Steel Indexing Plungers

GN 617.1
Indexing Plungers with Rest Position, Stainless Steel / Plastic Knob

GN 617.1
Stainless Steel Indexing Plungers

GN 613
Indexing Plungers, Steel / Plastic Knob

(b) Illustration of number range search and filtering on *indexing plungers* from website of spare part manufacturer [29].

Figure 14. Illustration of filtering options, facets and number range search on the website of the standard spare part manufacturer.

ported by the search. This is verified by searching for multiple spare parts by semantic descriptions. Therefore, search by means of synonyms or semantic descriptions of spare parts, e.g. *round*, *red*, *open/close*, etc., is not feasible. For convenience, the search engine deployed for comparison will be called *text search* in the following.

4.1.5 Test Design

The goal of this thesis is the evaluation of the performance of visual search in comparison to text-based search combined with navigation. Therefore, a **comparative summative test** is chosen. In comparative tests, the same users can attempt tasks on all search engines (within-subjects design) or different sets of users can work with each product (between-subjects design) [22]. Within the scope of the present thesis, a **within-subjects design** is selected since this design has **greater statistical power** than a between-subjects design. This is beneficial since fewer participants are needed in order to find statistically significant effects, as each participant is evaluating more than one design at a time (and thus each version is being evaluated in more instances).[23]

Test Setup

The evaluation experiment is conducted by means of unmoderated remote test sessions on the software *Loop11*. Participants are given 16 images of spare parts and asked to search for these spare parts on the two search systems under evaluation. The search tasks are given one-by-one and the search systems, i.e. visual and text-based search are alternating. Two structurally identical test setups with two different sets of search tasks are assembled for the evaluation experiment. The configuration of the test sets is given in Table 3.

Table 3. Within-subjects test design.

Search System	**Test Setup A**	**Test Setup B**
Text-based search	Part No. 1-8	Part No. 9-16
Visual search	Part No. 9-16	Part No. 1-8

The two different sets of search task are assembled from the same pool of 16 spare parts to ensure that all spare parts are searched through both search engines. The test procedure is based on alternating text-based and visual search tasks to keep the participants engaged and prevent them from losing motivation. Participants never perform the same task twice across the two search systems. This helps to reduce the learning effect for how they think of and approach the search process.[23] The attempt of reducing the learning effect is based on the assumption that since the sample size is small the learning effect is hard to measure. Thus, the aim is to avoid the learning effect as an additional variable in the experimental setup.

Testing Procedure

After registration of the candidate, the qualified test participant receives a short video with instructions and the link to the respective test set. The software deployed for the test, Loop11, provides an online software environment that allows the creation of a survey that includes tasks for users to complete on an external website. For presentation purposes, the testing procedure for test set A is presented in the Table 4.

Table 4. Testing procedure.

Task Type	**Task Description**
Introduction	
Task 1	Part No. 1-8
Task 1	Download of image of spare part No. 1 (Semi-Split Shaft Collars)
Task 2	Search for spare part No. 1 in text search
Question 1	Input of Query Result
Task 3	Download of image spare part No. 9 (Adjustable Hand Levers)
Task 4	Search for spare part No. 9 in visual search
Question 2	Input of Query Result
Task 5	Download of spare part No. 2 (Stainless Steel Serrated Locking plates)
Task 6	Search for spare part No. 2 in text search
...	...
Question 49	Rating of search experience

The software allows for the measurement of **task time**. The task time is, as defined in chapter 3.1.2, the total duration of time that a user spends on searching the spare part. It starts when the user clicks on *Start Search Task* and ends when the user clicks on *Task Completed*. Subsequently, the user is asked to input the search query result as keywords. After the completion of the search test, the participants are asked to respond to a **post-test questionnaire** to rate their search experience. The post-test questionnaire is based on rating scale items that are characterized by closed-ended response options [22]. Participants are asked to agree or disagree to a statement based on a rating scale. The analysis of the resulting time metrics and success rates, as well as a summary of the results from the post-test questionnaire are defined in the following chapter.

4.2 Analysis of empirical metrics

81 observations of **task times** and **search query responses** are collected for a total number of 16 tasks and 2 test sets. Thus, a data set with dimensionality of 5184 values is obtained. The data is distributed over eight distinct data sets (see Chapter 4.1.5). One data set per metric and per test setup for each sample group, i.e. expert or support, as illustrated in Table 5. First, the data set is pre-processed and flawed observations are removed.

Table 5. Obtained data sets after pre-processing.

	Expert Sample		**Support Sample**	
	Test Setup A	**Test Setup B**	**Test Setup A**	**Test Setup B**
Task Time	Text (1-8) Visual (9-16)	Text (9-16) Visual (1-8)	Text (1-8) Visual (9-16)	Text (9-16) Visual (1-8)
Search Success	Text (1-8) Visual (9-16)	Text (9-16) Visual (1-8)	Text (1-8) Visual (9-16)	Text (9-16) Visual (1-8)

4.2.1 Treatment of Missing Values

After data collection is completed, 27 flawed observations are identified and removed from the data set. The flawed test observations have their origin in technical limitations of the testing environment, such as bad internet connections or add-on blockers. Additionally, entries that have been conducted on mobile application instead of computers are removed since the search experience in the given test environment is completely different and therefore statistically not comparable. The aforementioned removal of data sets is controlled by removing only those entries with a minimum amount of 16 missing questions or tasks. The **resulting data set consists of** 54 **observations**.

During further analysis, missing values of the task time are omitted in the analysis. Therefore, the missing values are evaded by adjusting the population size for the respective task or participant respectively. Missing values of search success are not omitted in the analysis but classified as FALSE since the search is defined as unsuccessful.

4.2.2 Task Time and Success Rate

Task time is a *continuous measurement* that is obtained in ISO 8601 format, i.e. *hh:mm:ss*, and subsequently converted into seconds. **Search success** is collected as a *binary measure* of search task success or task failure. The binary measure takes the value of TRUE (or 1) if a spare part is identified correctly and a value of FALSE if the spare part is not identified correctly. A spare part is classified as correctly identified if the query result response by the participant contains the correct SKU, part ID, group ID or parts of the full name that are unique to the respective spare part. Also classified as correctly identified are results that belong to parts that fulfill the **exact same functionality but have different material properties**. This decision is based on the assumption that a participant may assess the material better when facing the spare part in reality than when looking at an image of the part.

4.2.3 Evaluation of Differences

The mean (i.e. average) of task time and success rate are computed per task and per participant. Based on these averages, the difference in task time, and success rate respectively, are calculated and their estimates of variability determined. The difference metrics with the smaller variance, i.e. the difference per participant, is chosen for further statistical analysis. The **difference per participant in task time** ($\Delta(t_k)$) is calculated between average task time on text search tasks and average task time on visual search tasks for each of the four data sets (i.e. for both sample groups and both test sets) according to the following formula:

$$\Delta(t_k) = \frac{(t_{text1,k} + t_{text2,k} + ... + t_{text8,k})}{n} - \frac{(t_{visual9,k} + t_{visual10,k} \cdots + t_{visual16,k})}{n} \tag{1}$$

where $t_{texti,k}$ is the time taken by participant k on the text search task for spare part no. i and $t_{visualj,k}$ is the time taken on the visual search task for part no. j. n is the number of task time measurements in each test set per search system, i.e. missing values of time taken are omitted.

Similarly, the **difference per participant in success rate** is a continuous measurement that is calculated for each of the four data sets according to the following formula:

$$\Delta(sr_k) = \frac{s_{visual1,k} + s_{visual2,k} + ... + s_{visual8,k}}{n} - \frac{s_{text1,k} + s_{text2,k} + ... + s_{text8,k}}{n} \tag{2}$$

where $s_{texti,k}$ is the binary success measure of participant k on the text search task for spare part no. i and $s_{visualj,k}$ is the success measure on the visual search task for part no. j. n is the number of search tasks in each test set per search system.

4.2.4 Descriptive Statistics

The difference metrics per participant, i.e. task time and search success, are analysed in terms of descriptive statistics. To this end, estimates of location (e.g. mean, outliers), estimates of variability (e.g. variance, standard deviation) and the data distribution are explored. The average metrics for the participants are used to identify outliers among participants. Outliers are data values that are very different from most of the data [30]. The within-subjects test design removes a major source of variation between sets of data, i.e. the variation between users. However, the estimates of variability are critical to evaluate since small variance in data are crucial in observing any statistical differences between versions of a design [23]. The data distribution is explored by means of box plots and histograms to test for normality and outliers.

4.2.5 Treatment of Outliers

The estimates of variability are especially sensitive to outliers since they are based on the squared deviations [30]. Outlier analysis is crucial since the estimates of variability will be deployed in significance testing later on. The difference metrics are examined for outliers by means of box plots and the reasons for the existence of detected outliers are analysed. Additionally, the outliers are examined regarding the likelihood that similar values will continue to appear. The outliers are detected by inspecting the differences and singling out the corresponding observation. Subsequently, the screen recording of the respective observation is reviewed. The screen recordings are scanned to rule out the possibility of technical problems and negligent testing behaviour. In three cases negligent testing behaviour is detected as participants ignore the visual search and only apply text search. A reason for this might be the fact that the images of the spare parts need to be downloaded in order to apply the visual search. Therefore, the **final data set for analysis consists of** 51 **observations**. Of those, 21 belong to the expert sample and 30 to the support sample.

4.3 Validation of Metrics: Statistical Significance Testing

To assess a difference in the performance of visual search compared to text-based search for the spare part industry as a whole, the statistical significance of the differences for the samples must be assessed. Therefore, the metrics of visual and text search are tested on whether their difference is greater than what would be expected from chance, i.e. determining whether the observed effect lies within the range of normal chance variation. Thereby, the size of the difference that might likely exist in the untested population is assessed.[22] Every significance test, also called hypothesis test, is based on a null hypothesis [9].

The null hypothesis states that any effect that is observed is due to random chance. The alternative hypothesis is the counterpoint to the null hypothesis.[30] In this case, the **alternative hypothesis states that the task time is shorter for visual search and the success rate is higher for visual search**, respectively. In within-subjects test designs, all participants conduct search tasks on both text and visual search. Therefore, a paired t-test can be applied to assess the results.[22]

Paired t-Test

A paired (or *dependent*) t-test is used when the observations are not independent of one another. Since in the within-subjects test design the participants conduct tasks on both text search and visual search, a relationship between the metrics is expected. The paired t-test is applied to the differences in the values of the two variables and tests if the mean of these differences is equal to zero.[27] According to Albert et al., the paired t-test can also be called a difference score t-test because it is based on the mean and standard deviation of the difference scores rather than the raw scores. The paired t-test is based on the following assumptions:

- independence: measurements for one participant do not influence measurements for another participant.
- each pair of measurement must be obtained from the same participant: visual and text-based measurements
- measured differences are normally distributed
- no extreme outliers: the differences should not contain any extreme outliers.

In the following the procedure of the paired t-test is presented according to Croft et al.[9]. First, the metric in question, i.e. average task time or average success rate, is computed for both search systems. The difference between the metric obtained through visual and text search is computed and a hypothesis is made. Subsequently, the test statistic (also t-statistic) is computed based on the mean difference $\hat{d}$, the standard deviation sd of the difference and the number of samples n according to the following formula:

$$t = \frac{\hat{d}}{\frac{sd}{\sqrt{n}}} \tag{3}$$

The t-value stands for the calculated difference represented in units of standard error. The greater the magnitude of the t-value, the greater the evidence against the null hypothesis. The t-value is used to compute a p-value, which is the probability that a significant difference could be observed if the null hypothesis were true, i.e. if any observed effect is due to random chance. The null hypothesis (no difference) is rejected in favor of the alternate hypothesis if the p-value is smaller than the significance level α.[9] α is the probability threshold that chance results must surpass for actual outcomes to be deemed statistically significant [30]. In other words, the significance level of a statistical test is the probability that the test outcome could have occurred by chance (i.e. false positive conclusion). For this analysis, α is selected to be 0.05. This means that the results only have a 5 % chance of occurring, or less, if the null hypothesis is actually true. According to Sauro et al., although sample mean task times will differ from their population median, the paired t-test can still accurately tell whether the difference between means is greater than what would be expected from chance alone [22].

5 Empirical Results of Benchmark Test and Analysis

The results of the quantitative usability testing are presented and their statistical significance is analysed. In the first section, the results of pre-processing and the descriptive analysis are given. An overview of the data sets is displayed and examined. Thereby, the difference metrics for which the hypothesis holds are analysed. In the second section, the results of the statistical significance testing of the difference in task time and success rate between text and visual search are presented. The results of the qualitative post-test questionnaires are summarised in the third section.

5.1 Results of Descriptive Statistics

51 participants spend on average around 50 minutes on the test. In the expert sample, 4.8 % of search tasks are abandoned, i.e. no query results response is submitted. In the support sample, 12.7 % responses are left blank. Figure 15 and Figure 16 illustrate the **average time per task** for all test data sets. Although the scales differ, the distributions of task time across tasks from the expert and support sample resemble each other. Part no. 8 and no. 9 took participants the longest to find by means of text search. In case of the visual search, participants needed most time for the search parts no. 9 and no. 1.

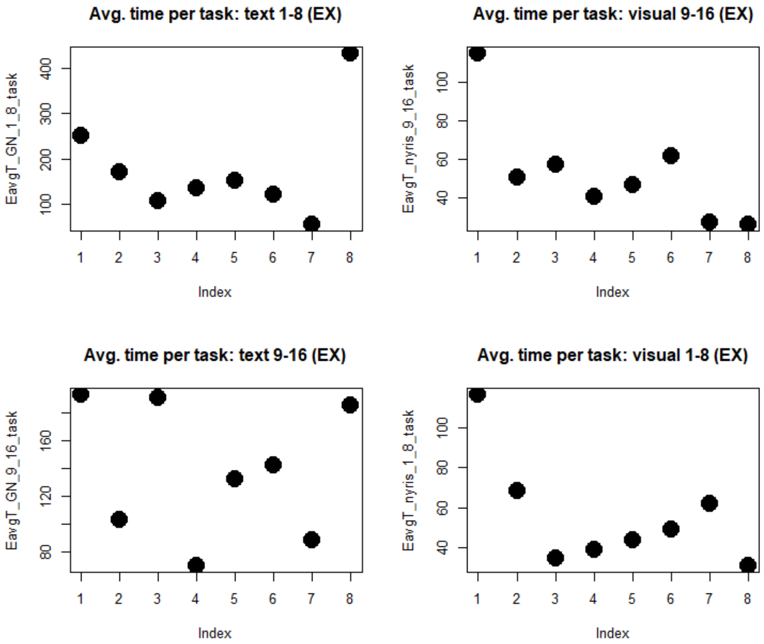


Figure 15. Average time per task (expert group) [s].

The **success rate per task** is illustrated in Figure 17 and Figure 18 for the expert and support

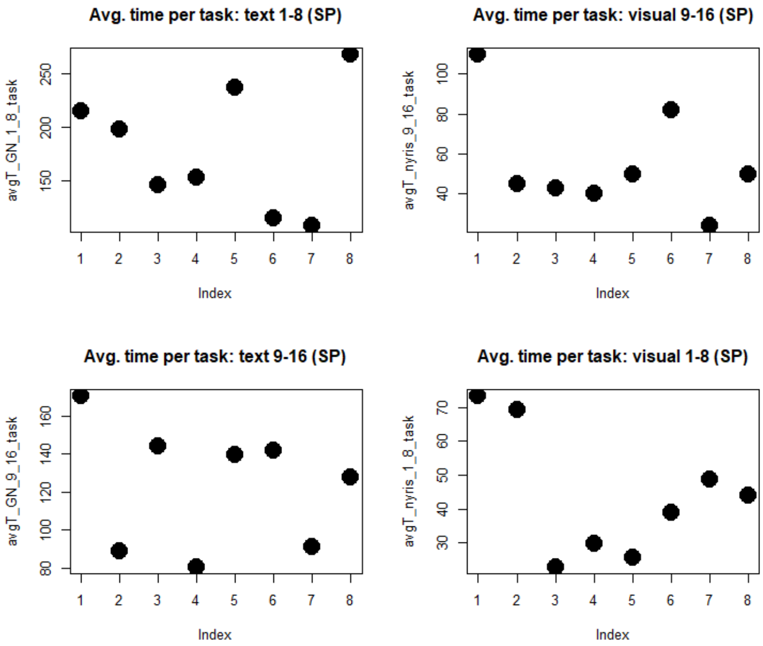


Figure 16. Average time per task (support group) [s].

sample, respectively. In contrast to the task time, the distribution of success rates differs quite significantly between the expert and the support sample observations. The expert sample has the lowest success rates for text search for parts no. 5, no. 8 and no. 11 and the support sample for parts no. 8 and no. 14. When applying visual search both sample groups had the greatest difficulty in finding parts no. 7 and no. 14. Since the variance across participants is smaller than the variance across tasks, the difference is analysed on the basis of the averages per participant.

5.1.1 Estimates of Location and Variability

Multiple estimates of location and variability of the **differences in task time** between visual and text search are described in Table 6. **Visual search appears to be faster across all test sets and sample groups**. However, the task time difference is higher for test set A than for test set B, both for the expert and the support sample. Similarly, the standard deviation in task time is higher for test set A. Overall, the performance between the expert and the support sample is similar with regard to the mean difference in task time. The minimum value of the difference in test set A of the support sample, i.e. -15.18, stands out since the mean task time for at least one participant when performing text search is lower than when performing visual search.

The estimates of location and variability of the **differences in success rate** between visual and text search are described in Table 7.Considering that a success rate of one represents the complete

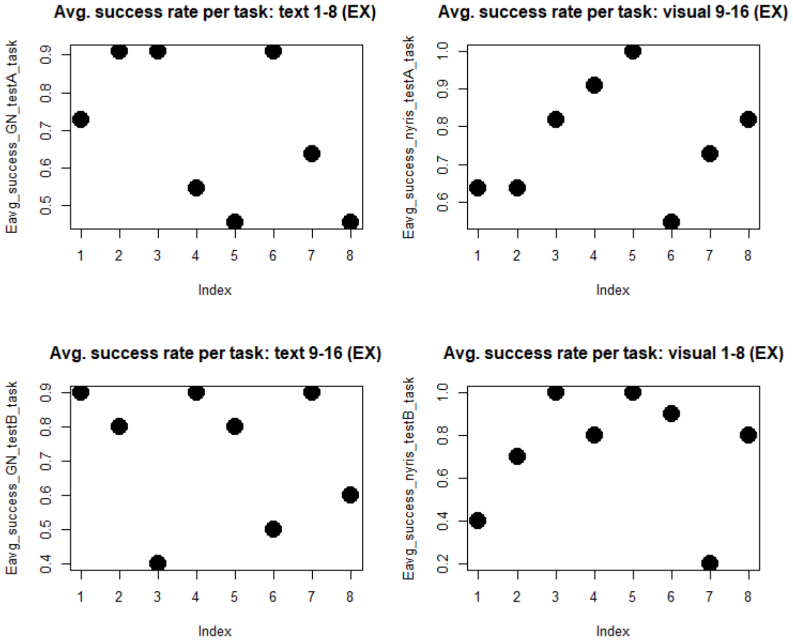


Figure 17. Average success rate per task (expert group).

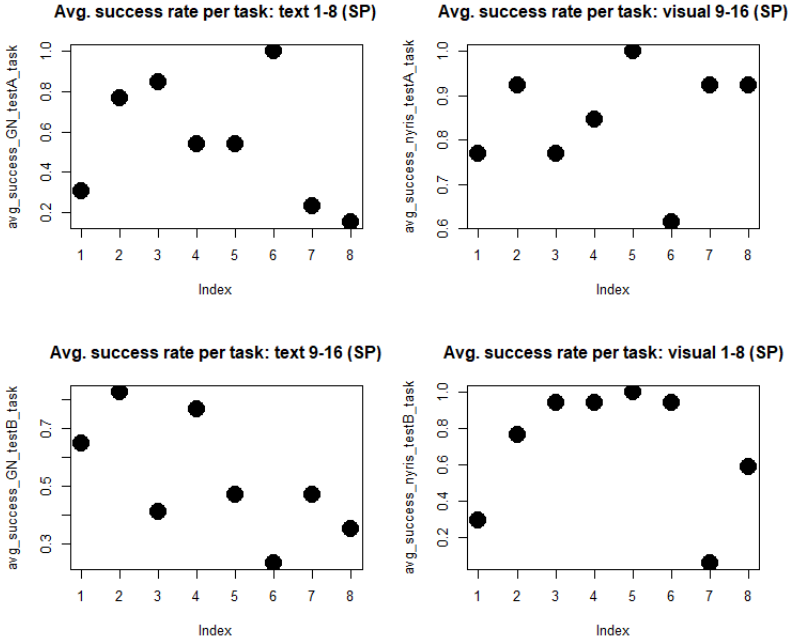


Figure 18. Average success rate per task (support group).

search success across all tasks, the mean **difference in success rate** between visual and text search appears to be very small for the expert sample but evidently higher in the support sample. The standard deviation of the difference in success rate for test set A is almost twice as high for the expert sample as for the support sample. However, for test set B, the contrary is the case. Across all sample groups and test sets, observations with a negative difference appear as well, i.e. text search was for these observations more successful than visual search.

Table 6. Descriptive statistics of resulting task time difference.

	Expert Sample		**Support Sample**	
	Test Set A	**Test Set B**	**Test Set A**	**Test Set B**
Sample Size	11	10	13	17
Mean [s]	124.79	82.59	124.56	78.95
Standard Deviation [s]	54.21	46.64	66.54	39.12
Maximum [s]	218	178.63	235.25	141.88
Minimum [s]	42.38	23.75	−15.18	3.25

Table 7. Descriptive statistics of resulting success rate difference.

	Expert Sample		**Support Sample**	
	Test Set A	**Test Set B**	**Test Set A**	**Test Set B**
Sample Size	11	10	13	17
Mean	0.07	0	0.30	0.17
Standard Deviation	0.29	0.20	0.15	0.30
Maximum	0.5	0.375	0.5	0.625
Minimum	−0.5	−0.25	−0.13	−0.25

5.1.2 Outliers

As elaborated in Section 4.2.5, outliers are crucial in determining the significance of the difference between visual and text search. The presence of outliers may decrease normality and therefore increase the error variance, thereby reducing the power of the paired t-test.[11] After the removal of flawed observations, the box plots of the difference metrics, depicted in Figure 19 and Figure 20, are examined for outliers.

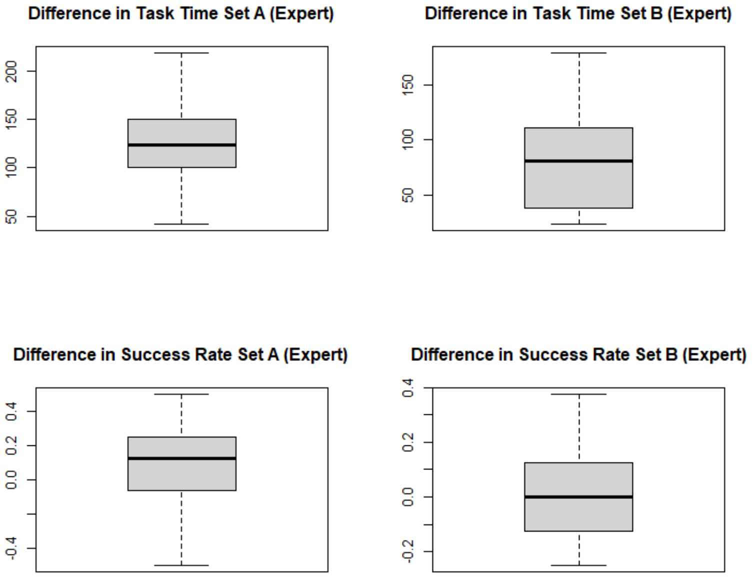


Figure 19. Box plot of differences in task time [s] and success rate for both test sets A and B (expert group).

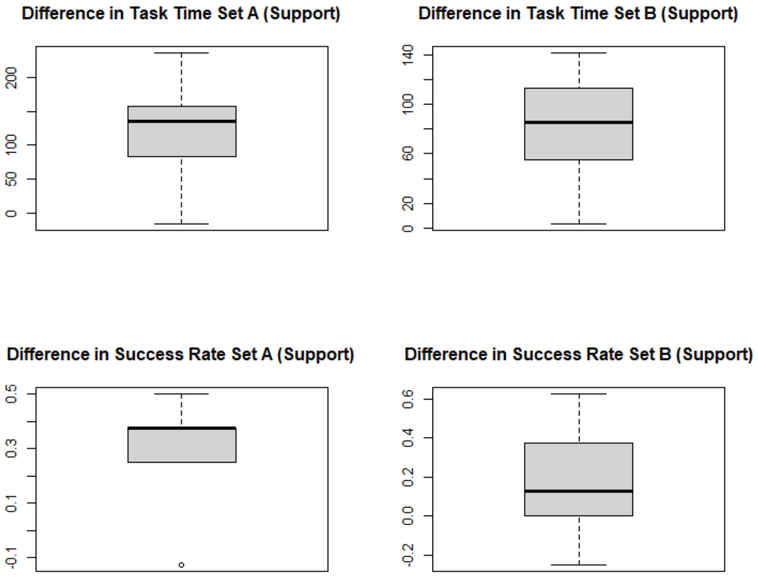


Figure 20. Box plot of differences in task time [s] and success rate for both test sets A and B (support group).

Outliers are only observed in the mean difference in success rate for test set A in the support sample (see Figure 20). All other difference metrics appear to not contain any outliers according to the box plot diagrams .

5.1.3 Normality

The distributions of differences in task time and success rate for both test sets and both sample groups are presented in Figure 21 and Figure 22. The distributions of differences tend towards a normal distribution for almost all test sets. However, the distribution of the difference in success rate for test set A in the support sample displays non-normality. This correlates to the result of the outlier analysis since test set A of the support sample contains at least one outlier. However, since the majority of differences tends towards normality, the paired t-test is applied to test for statistical significance in the following.

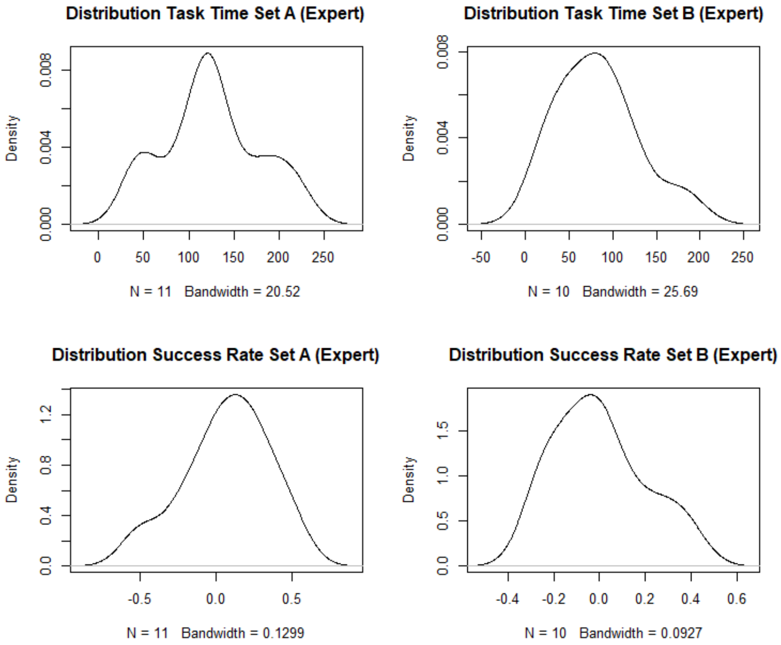


Figure 21. Distribution of differences in task time [s] and success rate for both test sets A and B (expert group).

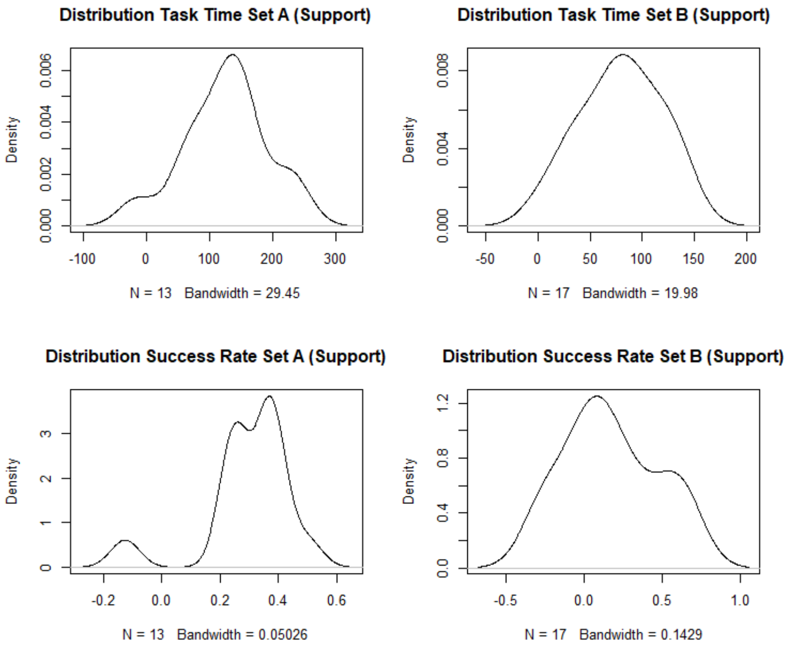


Figure 22. Distribution of differences in task time [s] and success rate for both test sets A and B (support group).

5.2 Results of Statistical Significance Testing

The results of the paired t-test for all differences in task time and success rate for both test sets and sample groups are described in Table 8 and in Table 10. As elaborated in Chapter 4.3, the p-value is the probability that a significant difference could be observed if the null hypothesis were true, i.e. if any observed effect is due to random chance. If the p-value is smaller than chosen the significance level $\alpha = 0.05$, i.e. the probability that the test outcome could have occurred by chance, then the respective difference may be declared as statistically significant.

5.2.1 Task Time

Since the p-values of the applied paired t-tests are smaller than the significance level, $\alpha = 0.05$, the mean difference in task time is statistically significant for both the expert and the support sample group on both test sets. Therefore, **visual search is statistically significantly faster than text search**. The mean difference in search task time ranges from 78.95 s for the support sample on test set B to 124.79 s for the expert sample on test A. The confidence interval for the mean specifies a range of values within which the unknown population parameter, in this case the mean difference in task time between visual and text search, may lie.[27] The confidence intervals for the differences in task time are presented in Table 9.

Table 8. Results of paired t-test for difference in task time.

	Expert Sample		**Support Sample**	
	Test Set A	**Test Set B**	**Test Set A**	**Test Set B**
Sample Size	11	10	13	17
Mean [s]	124.79	82.59	124.56	78.95
Standard Deviation [s]	54.21	46.64	66.54	39.12
t-value	7.64	5.60	6.75	8.32
p-value	1.77×10^{-5}	3.34×10^{-5}	2.05×10^{-5}	3.32×10^{-7}

Table 9. Confidence intervals of differences in task time.

	Expert Sample		**Support Sample**	
	Test Set A	**Test Set B**	**Test Set A**	**Test Set B**
Sample Size	11	10	13	17
Task Time [s]	88.38 – 161.21	49.23 – 115.95	84.36 – 164.77	58.83 – 99.06

5.2.2 Success Rate

The results of the paired t-test for the difference in success rate between visual and text search, described in Table 10, differ significantly from the task time results. For the expert sample, the p-values for both test set A and test set B are much greater than the significance level α. This means that the null hypothesis can not be rejected and the difference in mean for test set A, i.e. 0.07, could have occurred by chance. However, for the support sample, there appears to be a statistically significant mean difference in success rate of 0.30 for test set A and 0.17 for test set B between visual and text search since the p-values for both test sets are smaller than the significance level. The confidence intervals for the differences in success rate are presented in Table 11.

Table 10. Results of paired t-test for difference in success rate.

	Expert Sample		**Support Sample**	
	Test Set A	**Test Set B**	**Test Set A**	**Test Set B**
Sample Size	11	10	13	17
Mean	0.07	0	0.30	0.17
Standard Deviation	0.29	0.20	0.15	0.30
t-value	0.79	0	7.21	2.33
p-value	0.45	1	1.08×10^{-5}	0.03

Table 11. Confidence intervals of differences in success rate.

	Expert Sample		**Support Sample**	
	Test Set A	**Test Set B**	**Test Set A**	**Test Set B**
Sample Size	11	10	13	17
Success Rate	$-0.12 - 0.26$	$-0.15 - 0.15$	$0.21 - 0.39$	$0.02 - 0.32$

5.3 Results of Qualitative Satisfaction Questionnaires

In Table 12 the results of the post-test questionnaires that participants receive at the end of a usability session are presented. The responses of all participants are taken into account, including those who are removed from the quantitative analysis. An accumulated 89.1 % of participants perceive the search experience with visual search to be better than with text search and navigation, while only 3.6 % feel that text search is better. An accumulated amount of 92.7 % think that the results returned by visual search are relevant. In contrast, only 25.4 % of participants think that the results returned by text search are relevant. While 85.4 % of participants would like to visually search in their private life, 94.6 % would like to apply visual search in their job. Considering that 41.2 % of participants belong to the expert sample, this means that the **vast majority of maintenance technicians and spare parts managers would like to apply visual search on the job**.

Table 12. Results of post-test rating questionnaire.

Answers from 55 participants	**Strongly agree**	**Agree**	**Neither agree nor disagree**	**Disagree**	**Strongly disagree**
The search experience was better with visual search.	36 65.5 %	13 23.6 %	5 9.1 %	1 1.8 %	0 0.0 %
The search experience was better with text search.	0 0.0 %	2 3.6 %	14 25.5 %	22 40.0 %	17 30.9 %
The search results of the visual search were relevant.	24 43.6 %	27 49.1 %	3 5.5 %	1 1.8 %	0 0.0 %
The search results of the text search were relevant.	2 3.6 %	12 21.8 %	25 45.5 %	15 27.3 %	1 1.8 %
I would like to use visual search on my job.	38 69.1 %	14 25.5 %	2 3.6 %	1 1.8 %	0 0.0 %
I would like to use visual search in my private life.	34 61.8 %	13 23.6 %	1 1.8 %	5 9.1 %	2 3.6 %

6 Discussion and Conclusions

This thesis tests the hypothesis that visual search is more successful and faster than text-based search and navigation. In this chapter, the results are discussed and the limitations of the thesis are critically reflected upon. Based on this, areas with the need for further research are identified.

6.1 Critical Reflection

This thesis asserts that the null hypothesis can be rejected for the time difference of search tasks, i.e. the search time is significantly shorter for visual than for text search. For the success rate, the null hypothesis can be rejected for one test set. Thus, the success rate can be interpreted as partly higher for laypeople, while it does not enhance the success rate for maintenance technicians and spare parts managers. Across both samples, participants perceive the visual search as better (see Chapter 5.3). However, there are multiple arguments that imply the acceptance of the null hypothesis, i.e. a significant difference does not exist, and the rejection of the alternative hypothesis. These will be elaborated on in the following sections.

6.1.1 Bias in Participant Sampling

The most important concepts in participant sampling are representativeness and randomness [22]. The representativeness of the expert group sample for the population of maintenance technicians and spare part managers is ensured by their qualification, a further factor being their indication in advance of the average amount of time that they are spending on spare part search. For the support sample group, the representativeness is difficult to assess. Since convenience and snowball sampling is applied, participants from the expert group are likely to have similar background knowledge of spare parts. Likewise, participants from the support group might share a similar approach at deploying on-site navigation. Therefore, complete **randomness cannot be ensured** for neither sample group.

6.1.2 Bias in Test Environment

The aim of this thesis is to test the hypothesis that visual search is more efficient than text search for the population of maintenance technicians and spare part managers. However, in general, spare part search is not conducted on the basis of a photo of the spare part. In most cases, the maintenance technician dismounts the spare part and has it on hand. Thereby, the technician can judge the material and finishing of the spare part. The test design therefore creates an imbalance and bias that favours the visual search, i.e. the RIS technology. Within the scope of this thesis, this bias is countered by classifying those search responses which contain parts with the same functionality as successfully found (i.e. TRUE). Even when their material or finishing differs from the spare part provided in the query image.

6.1.3 Test Design

Within the scope of the thesis at hand, a within-subjects usability test design is chosen since it has greater statistical power for small sample sizes as the variance is partitioned. However, for the within-subjects test design and the assembled test setup, the **influence of task effects** cannot be excluded. Participants may be faster, for instance, at a specific text search task since they know which keyword to enter from a previous search task in the same test set. The influence of sequence is minimised by splitting up similar spare part queries across the visual and the text search. The optimal test setup in order to decrease the sequence effect is a completely random selection of tasks from the test corpus for each participant. However, this is not feasible with the software deployed to run the test. The test sets are assembled in such a way that participants conduct text and visual search tasks in alternating order. This might reduce the learning effect. However, since 16 parts are searched by each participant, there is a risk that participants grow weary of the text search and abandon tasks quicker.

In summary, although the within-subjects design excludes the greatest origin of variance, i.e. variance among participants, it contains many factors, such as the task effects, that add to bias and variance. Therefore, a between-subjects design is favourable if a sufficiently large sample can be obtained. In a between-subjects design, each participant interacts only with one test system. Thus, the between-subjects design removes any bias regarding the perception of the users of the two search engines. In a between-subjects design any statistically significant results are attributed solely to the difference in the systems being tested [23]. Therefore, a future test deploying a **between-subjects test design might be valuable to further add to the findings of the present study**.

6.1.4 Search Task Abundance

As elaborated in Section 5.1, 4.8 % in the expert sample and 12.7 % of search query responses in the support sample are left blank. In absolute numbers these are 77 abandoned tasks out of a total of 816 search tasks. For the purpose of analysis, these missing responses are classified as FALSE, i.e. the spare part search is not successful. Search abundance may, however, originate from two different reasons. Either the spare part can indeed not be found or the search is abandoned due to long search times. This differentiation is not taken into consideration within the scope of this thesis. The idea of combining the search abundance with the search time leads back to the metric of efficiency introduced in section 3.1.2. Thus, the incorporation of an efficiency metric might add to the significance of the obtained findings of the present thesis.

6.1.5 Statistical Significance of Results

The statistical significance for these metrics is determined by applying a paired t-test, resulting in the p-value, i.e. the probability that, **given a chance model**, differences as extreme as the observed may occur. However, according to the American Statistical Association (ASA), the p-value does not measure the probability that the studied hypothesis is true [31]. Neither does the p-value measure the size of an effect or the importance of a result [31], nor is a hypothesis that can not be rejected automatically true. The reason for this is that the decision on statistical significance in accordance with the paired t-test is based on a given chance model [30].

The chance model of the paired t-test is based on the normality assumption for the differences, and supposes that these differences do not contain any outlier. Therefore, the results from the paired t-test must be examined taking the distributions and outliers into account. If the differences are statistically significant, that means they are very unlikely to occur if the null hypothesis is true. Hence, the null hypothesis is rejected. This could, however, actually be a type I error (false positive). If the difference however does not show any statistical significance, it has a high chance of occurring if the null hypothesis is true. Therefore, the null hypothesis fails to get rejected. This however may also be in fact a type II error (false negative).[30]

Although the difference in success rate for test set A in the support sample has a p-value of $1.08e-5$ and the highest mean difference across all test sets, the box plot of the difference reveals an outlier (see Figure 20. The outlier participant has a higher success rate when applying text search than when using the visual search application. This outlier leads to the distribution of differences in success rate being clearly non-normal. Therefore, the statistical significance in the difference of success rate for test set A in the support group might actually be a type I error, i.e. the null hypothesis is rejected when it is in fact true. To test for statistical significance in this case of non-normality a non-parametric test may be more appropriate. However, it is likely that the non-normality comes from the small sample size. In fact, the distributions across all test sets and metrics are not perfectly normal but slightly skewed. Therefore, a larger sample size in general will likely lead to more normal distributions to which the paired t-test could be applied more successfully.

6.2 Practical Significance

Despite the arguments in Section 6.1, challenging the findings of this thesis, the application of visual search is evidently faster and perceived as more favourable by users. Statistically significant differences in task time are observed across all test sets, while a significant difference in success rate is detected for one test set of the support group. In practical terms this means that while laypeople experience higher success rates and shorter search times than by traditional text search, experts find spare parts at similar success rates faster. This raises the hypothesis that visual search can enable laypeople to find spare parts as accurately as experts by means of visual search. The 95 % confidence intervals of time difference ranges from 49 s to 161 s for maintenance technicians and from 59 s to 165 s for laypeople. These time ranges are likely to be perceived by the search engine user. The confidence intervals for the differences are described for all test sets in Table 13. Therefore, assuming a downtime cost of 142 €/s, visual search can save large manufacturing companies in between 6958 € and 22 862 € per spare part search [1].

Table 13. Confidence intervals of differences.

	Expert Sample		Support Sample	
	Test Set A	Test Set B	Test Set A	Test Set B
Sample Size	11	10	13	17
Task Time [s]	88.38 – 161.21	49.23 – 115.95	84.36 – 164.77	58.83 – 99.06
Success Rate	n.a.	n.a.	n.a.	0.02 – 0.32

6.3 Domains for Further Research

The obtained statistically significant findings on differences in task time and success rate between visual and text search, as well as the perceived usability of visual search, can be deployed as benchmark for further research. The following three hypotheses arise from this thesis:

1. A statistically significant difference between success rates of visual and text search is found with a larger sample size.
2. The asserted hypothesis of a statistically significant difference in task time is fortified by the conduction of a between-subjects test design.
3. Visual search can enable laypeople to find spare parts as accurately as experts by means of visual search.

7 Summary

The main goal of this thesis is testing the hypothesis that visual search is faster and more accurate than text search and navigation in spare part search. The hypothesis is tested by means of a within-subjects experiment on two search engines, a visual search engine from *nyris*, a software development company from Berlin (Germany), and a text-based e-commerce search on the website of a standard spare part manufacturer. The evaluation corpus is based on the products and services catalogue from a standard spare part manufacturer. The data set consists of 34 534 standard spare parts with a total number of 34 817 images. The experiment is conducted by means of unmoderated remote test sessions on the software *Loop11*. The test sessions are conducted with 54 participants. The participants are split into the following two sample groups:

- expert sample: 23 qualified maintenance technicians and spare parts managers,
- support sample: 30 mechanical engineers and laypeople.

The maintenance technicians and spare parts managers come from 11 different companies across multiple industries, who **spend an average of around 8 hours on spare part search each week**, see Figure 11.

Participants are given a test set of 16 randomly selected standard spare parts and asked to search for these spare parts on the two search systems. The search tasks are given one-by-one and the search systems, i.e. visual and text search, are alternating. Two structurally identical test setups with two different sets of search tasks are assembled for the evaluation experiment. The configuration of the test sets is given in Table 3. In the experiment, the **search task time** and the **search success** are measured. The test procedure is concluded with a **post-test questionnaire**, where participants are asked to rate their search experience **qualitatively**.

At first glance, differences between visual and text search are evident across both task time and success rate in multiple test sets. **Visual search is significantly faster than text search** for both sample groups and **more successful for the support sample**. The mean difference in search task time ranges from 78.95 s for the support sample on test set B to 124.79 s for the expert sample on test A. Statistical significance tests are conducted by means of the paired t-test. Since for the difference in task time the p-values are smaller than the significance level, $\alpha = 0.05$, the **mean difference in task time is statistically significant for both the expert and the support sample group on both test sets**. For the difference in success rate, however, the p-values for both test sets in the expert sample are much greater than the significance level α. This means that the null hypothesis can not be rejected and the difference in mean for test set A, i.e. 0.07, could have occurred by chance. However, for the support sample, there appears to be a statistically significant mean difference in success rate of 0.30 in test set A and 0.17 in test set B since the p-values for both test sets are smaller than the significance level. Despite the appropriate p-value, the difference in success rate for test set A must, however, be rejected as **outliers** in the test set result in **non-normal behaviour**. In this case, the alternative hypothesis can not be accepted since the paired t-test is based on the assumption of normality.

The results of the post-test questionnaires demonstrate that despite the uncertainties in statistical significance, the visual search experience is perceived as favourable by users. An accumulated 89.1 % of participants think the search experience with visual search to be better than with text search and navigation, while only 3.6 % feel that text search is better. An accumulated 92.7 % think that the results returned by visual search are relevant. In contrast, only 25.4 % of participants think that the results returned by text search are relevant. 94.6 % would like to apply visual search in their job. Considering that 41.2 % of participants belong to the expert sample, this means that the **vast majority of maintenance technicians and spare parts managers would like to apply visual search on the job**.

In economic terms this means that for the observed confidence interval in time difference from 49 s to 161 s for maintenance technicians, and assuming a downtime cost of 142 € a second, visual search can save large manufacturing companies in between 6958 € and 22 862 € per spare part search [1]. Moreover, the findings of this thesis imply that in times of skilled labour shortage, **visual search might enable laypeople to take on functions in spare parts management**. Visual search therefore offers manufacturing companies the potential for savings and improvements that quickly pay for an investment in this search technology.

In conclusion, the combination of statistical results and the qualitative user perception substantiate the superiority of visual search in spare parts search. Even if the sample size is small and statistical uncertainty is observed, common sense allows the statement that visual search has **high potential in spare parts management**. While laypeople experience higher success rates and shorter search times than by traditional text search, experts find spare parts at similar success rates faster. The difference in success rates between laypeople and experts is unsurprising as it follows observation in daily life. Its reflection in the obtained results, strengthens the validity of this thesis.

The obtained findings can be deployed as benchmark for further research on the comparison of the usability of different search engines. In addition, this thesis raises the hypothesis that visual search can enable laypeople to find spare parts as accurately as experts. Further research can assess the statistical significance of the difference between success rates of visual and text search for a larger sample size. The asserted hypothesis of a statistically significant difference in task time could be fortified by the conduction of a between-subjects test design.

References

[1] "The True Cost of Downtime How much do leading manufacturers lose through inefficient maintenance?", [Online]. Available: www.senseye.io.

[2] *Playing Russian Roulette with Your Infrastructure Can Lead to Big Downtime - Aberdeen Strategy & Research*. [Online]. Available: https://www.aberdeen.com/techpro-essentials/playing-russian-roulette-with-your-infrastructure-can-lead-to-big-downtime/.

[3] H. Biedermann, *Ersatzteilmanagement: Effiziente Ersatzteillogistik für Industrieunternehmen*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, vol. 2, ISBN: 978-3-540-00850-7. DOI: 10.1007/978-3-540-68205-9. [Online]. Available: http://link.springer.com/10.1007/978-3-540-68205-9.

[4] G. Pawellek, *Integrierte Instandhaltung und Ersatzteillogistik*. Springer Berlin Heidelberg, 2016. DOI: 10.1007/978-3-662-48667-2.

[5] Roland Berger Strategy Consultants, "Online automotive parts sales: The rise of a new channel", 2014-05. [Online]. Available: https://www.rolandberger.com/publications/publication_pdf/roland_berger_study_online_automotive_parts_sales.pdf.

[6] C. Teixeira, I. Lopes, and M. Figueiredo, "Classification methodology for spare parts management combining maintenance and logistics perspectives", *Journal of Management Analytics*, vol. 5, no. 2, pp. 116–135, 2018-04, ISSN: 23270039. DOI: 10.1080/23270012.2018.1436989.

[7] C. Connolly, "Warehouse management technologies", DOI: 10.1108/02602280810856660. [Online]. Available: www.emeraldinsight.com/0260-2288.htm.

[8] B. Oumaima, A. E. Barkany, and A. E. Biyaali, "The performance evaluation of the spare parts management: Case study", *Management and Production Engineering Review*, vol. 10, no. 2, pp. 37–49, 2019. DOI: 10.24425/mper.2019.129567.

[9] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*, 2015 Update. Pearson, 2010.

[10] The Editors of Encyclopaedia Britannica, *Encyclopædia britannica: Search engine*, 2022. [Online]. Available: https://www.britannica.com/technology/search-engine.

[11] C. C. Aggarwal, "Data Mining: The Textbook", *Springer International Publishing*, pp. 285–344, 2015. DOI: 10.1007/978-3-319-14142-8{_}10.

[12] P. M. Chutel and A. Sakhare, "Evaluation of compact composite descriptor based reverse image search", 2017.

[13] M. Gaillard and E. Egyed-Zsigmond, "Large scale reverse image search a method comparison for almost identical image retrieval",

[14] R. Datta, D. Joshi, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age", *ACM Comput. Surv*, vol. 40, no. 5, p. 60, 2008. DOI: 10.1145/1348246.1348248. [Online]. Available: http://doi.acm.org/10.1145/.

[15] F. Pettersson and N. Pettersson, "Implementing an enterprise search platform using lucene.net", Ph.D. dissertation, Linköping University, 2012.

[16] R. Mukherjee and J. Mao, "Enterprise Search: Tough Stuff", *Queue*, vol. 2, no. 2, pp. 36–46, 2004-04, ISSN: 15427749. DOI: 10.1145/988392.988406. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/988392.988406.

[17] *How Algolia works | Algolia*. [Online]. Available: https://www.algolia.com/doc/guides/getting-started/how-algolia-works/#infrastructure.

[18] D. Tunkelang, “Faceted search”, *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 1, no. 1, pp. 1–80, 2009-01, ISSN: 1947-945X. DOI: 10.2200/s00190ed1v01y200904icr005.

[19] A. Oulasvirta and J. P. Hukkinen, “When more is less: The paradox of choice in search engine use”, in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, Association for Computing Machinery. Special Interest Group on Information Retrieval., 2009, pp. 516–523, ISBN: 9781605584836.

[20] *Iso 9241-11:2018(en), ergonomics of human-system interaction — part 11: Usability: Definitions and concepts*. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

[21] *Iso - iso/iec 25022:2016 - systems and software engineering — systems and software quality requirements and evaluation (square) — measurement of quality in use*. [Online]. Available: https://www.iso.org/standard/35746.html.

[22] J. Sauro, *Quantifying the user experience : practical statistics for user research.* Morgan Kaufmann, 2016, ISBN: 9780128023082.

[23] B. Albert, T. Tullis, and D. Tedesco, *Beyond the Usability Lab: Conducting Large-scale Online User Experience Studies*. Morgan Kaufmann Publishers, Elsevier, 2010.

[24] D. HAWKING DavidHawking, c. Nick Craswell, P. Bailey, and K. Griffiths, “Measuring search engine quality”, *Information Retrieval*, vol. 4, pp. 33–59, 2001.

[25] Y. Shang and L. Li, “Precision evaluation of search engines”,

[26] *ISO - ISO/IEC 25010:2011 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*. [Online]. Available: https://www.iso.org/standard/35733.html.

[27] *T-test | Stata Annotated Output*. [Online]. Available: https://stats.oarc.ucla.edu/stata/output/t-test/.

[28] *Spare parts search*. [Online]. Available: https://normparts.nyris.io/.

[29] *Ganter norm | home*. [Online]. Available: https://www.ganternorm.com/de/home.

[30] P. Bruce, A. Bruce, and P. Gedeck, *Practical Statistics for Data Scientists 50+ Essential Concepts Using R and Python*. 2016.

[31] R. L. Wasserstein and N. A. Lazar, “The ASA’s Statement on p-Values: Context, Process, and Purpose”, *American Statistician*, vol. 70, no. 2, pp. 129–133, 2016-04, ISSN: 15372731. DOI: 10.1080/00031305.2016.1154108/SUPPL{_}FILE/UTAS{_}A{_}1154108{_}SM8096.PDF. [Online]. Available: https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108.

Non-exclusive licence for reproduction and publication of a graduation thesis

I Friederike Mimi Freiin von Blomberg

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Evaluation of the Performance of a Visual Search Engine in Comparison to Text-based Search and Navigation”, supervised by René Pihlak and Daniel Beste.

1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons’ intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

16.05.2022

F.M. Blomberg