Radboud University Nijmegen

MASTER'S THESIS INFORMATION SCIENCE

A GENERIC DATA MODEL FOR STRATEGY MATCHING

8vance
manage your talent

Author:
Ahmed Taha

Supervisor RU:
prof. dr. ir. Th.P. (Theo) van der Weide

Supervisor RU:
dr. T.M. (Tom) Heskes

Supervisor 8vance:
dr. ir. J. (Jan) Jacobs

October, 2015

ACKNOWLEDGMENTS

During the period I have worked on the research leading to this paper, I received help from many people. Therefore, I would like to use this opportunity to extend my gratitude to everyone who has contributed to my research: I am in your debt.

First of all, I would like to thank 8vance founder Han Stoffels for giving me the opportunity to do this research for his company. Special thanks to my 8vance supervisor dr. Jan Jacobs for providing me with crucial advice and feedback during my research. The TIC stakeholder Floris van Eck has provided me with the necessary data, and this worked as an outstanding starting-point for my research. I would also like to thank all my other colleagues at 8vance and Océ for their support and hospitality. It was a pleasure working with all of you. The biggest gratitude has to be given to my supervisor professor Theo van der Weide for giving me the opportunity to work under his supervision, and to share his knowledge and experience with me as much as possible. I also would like to thank professor Tom Heskes for answering my questions and providing me with great feedback on my paper. The unconditional love and support of my family and friends during this research where absolutely vital and I am very grateful for it. Especially my girlfriend Bo who was always there for me.

Finally, I would like to dedicate this paper to my dearest friend and one of the most important people in my life, Ken Boogers ∞. Thank you for being an inspiration and an example for me, just by being who you are. I love you and will cherish every second we have spent together.

Eindhoven, October 2015

ABSTRACT

Relatively new trends like servitisation and big data are posing new opportunities & threats to the more traditional product based corporations. They are often confronted with a widening of the market space and consequently, a rapidly shrinking relevancy. For such corporate companies however, there is an opportunity to transform. Their current strategic course can be matched against startups within the same market with the goal to do strategic investments instead of short-term Mergers and Acquisitions (M&A).

The main goal of this study is to enable 8vance to take the first step in the strategy market using their Deep Matching Technology™. Therefore, this study is called Strategy Matching. The term Strategy Market implicates corporations that are interested in startup adoption with the goal to make a leap in their strategic course. For this study, the research questions are as follows:

- To what extent is it possible to establish a model suitable for the startup adoption market?
- What matching procedure is going to be used?
- What features are required for this model?
- How are these deliverables going to be validated?

As the first research question reveals, we tried to determine to what extent a generic model can be created which can function as an engine for the Strategy Matching market. This model can be seen as a theoretical description of the mathematical algorithms for the deep matching technique. The final theory proposal can be found in chapter six. Here, we explain that for matching, the main focus is similarity (and dissimilarity) and distance. Eventually, these worlds need to be connected using a mathematical solution.

In order to use the model properly, the data from both the startups and the strategic course of the interested corporation, need to be analysed and brought together in a unified framework. For this purpose, a set of features have been assembled. This feature set is divided into three classes: matching features, filtering features, and statistical features. Eventually, the matching features will be used to provide startups with the corresponding value. Also, the strategies can be quantified using these features. The next step for the corporation is to determine the order of importance of the matching features. This can be done using the AHP weighting approach described in chapter four. Then, the sub features need to be weighted and by integrating these results with the results from the feature weighting, the final matching result can be obtained. The main function of the filtering features is to focus on specific startup markets based on the selected strategy. The statistical features can provide additional, non-matching information (e.g. sentiment data) for the concerning corporation when considering to adopt a startup. In chapter five, the developed feature set and its AHP weighting are validated using TIC as a case study. For this process, the TIC stakeholder is asked to select a strategy and to provide the feature set with the corresponding weighting. Consequently, this is used to match a group of preselected startups which are then ranked on matching result. Finely, this ranking is compared with the intuitive ranking of the TIC stakeholder to show the quality of the developed Strategy Matching approach. The results where promising.

With this study, the first step into the strategy market space is taken. At this point, it is clear that further research is crucial to optimize the quality. The main points to focus on during further research are: automating the data gathering process since this is extremely time consuming, fine tuning the relative weighting of the features and sub features since this improves the quality of the matching results and looking into the additional features proposed by the TIC stakeholder. Also, what we would recommend looking into, is to find out whether the startup domain also is interested in adoption and whether they want to use this technique to draw attention of larger corporations. Lastly, it can be interesting to have a matching accuracy next to the matching result that indicates the extent of data obtained to complete the matching feature set.

CONTENT

1. INTRODUCTION

In the current business environment, company takeover is becoming increasingly popular. For larger corporations, it is often more interesting to adopt smaller companies rather than to invest time and money into gaining the same knowledge by themselves. With such acquisition, corporations can make a leap in implementing their strategic directions. But how do corporation interested in adoption, find the right startups? In other words, both the adopting company and startups need to be analysed. The next question would be, how do we compare the qualities of startups with the needs of a corporation based on its strategic course?

This means that a generic set of features need to be assembled which can be used to scout these startups. Then, these features need to be weighted to determine their hierarchical order. The next step is to gather startup data for comparison with corporation strategies. The gathered data can be compared using data mining methods in order to find the right match between startups and adopting corporations. The matching principle in this study will be further referred to as Strategy Matching. 8vance is a young Dutch software company that develops software systems to provide efficient data matching solutions which they call Deep Matching. Data mining algorithms are the key ingredient for deep matching. This study has been established as 8vance wants to examine the extent to which matching techniques can help in Strategy Matching.

In this study, we will set up the generic feature set and propose a data mining approach using similarity and distance algorithms. In chapter two, we discuss different similarity and distance measure techniques. We also discuss a technique that can be used for features weighting purposes. Chapter three contains the explicit requirements including the research questions. Based on these requirements, chapter four contains the feature set established during this study. An important step within the deep matching procedure is providing the matching feature set with their weighting based on the selected strategy. The weighting approach of the features and the determination of the final matching result is also described in chapter four. During this study, the matching feature set has been validated using 8vance's first launching costumer TIC, as an example of a corporation interested in startup adoption. Based on a selected strategy of TIC, a number of startups are subjected to a matching process, and the results, as well as the approach, are included in chapter five. Lastly, a detailed theory proposal is done in chapter six which addresses the various steps of data mining that we consider necessary in order to match the two mentioned parties. Note that for 8vance readers, emphasis is on chapters four and five. Chapter six is added for academic purposes.

2. LITERATURE

In this chapter we will list the literature that we initially have consulted during this study. First of all, similarity and dissimilarity is discussed. The Cosine Similarity and approach is discussed in section 2.3. In section 2.4, we have given a concise description of the Manhattan distance. In chapter six, this approach is discussed on a more abstract level. Lastly, we have added a general description of the AHP in the last section of this chapter. AHP is a technique we can use for weighting purposes. In section 4.6, a more detailed description is given including the mathematical approach within this study's context.

2.1 Matching challenge

When we speak of 'matching' in this context, we mean data from different parties that are brought together to review their similarities. [1] Coifman et al. indicate "working with the data in its original form can be quite difficult as the two sets typically consist of measurements of very different nature." This means that data primarily needs to be made measurable in order for it to be suitable for matching.

2.2 Similarity and Dissimilarity

About similarity and dissimilarity, [2] indicates "Objects are usually described as a set of properties (features). As a consequence, comparing two objects is done by comparing their properties. The result is called the similarity between two objects.

Similarity is a basic concept that is relevant in many kinds of applications. Similarity has some special properties. For example, we may find the sun and a volcano similar since they share 'fire', and we may find the sun and a ball similar since they share being spherical. Yet, from the similarity between sun and volcano, and the similarity between sun and ball, we may not conclude that volcano and ball also are similar. In technical terms, similarity is not a transitive relation between objects.

Furthermore, when we say that Peter is fighting like a lion, then we find that Peter (as an example) is similar to a lion (as a prototype for courage). But we may not conclude that a lion is like Peter. In technical terms, similarity is not a symmetric relation when comparing objects.

There are however some properties that are generally accepted as elementary requirements for similarity functions. Let X be a set of objects, and let $Sim : \mathsf{X} \times \mathsf{X} \rightarrow [0,1]$. First we focus on the (intended) meaning of the extremal values 0 and 1. We call objects x and y identical if $Sim(x, y) = 1$, and orthogonal when $Sim(x, y) = 0$: For convenience we introduce the following predicates:

- $Identical(x, y) \cong Sim(x, y) = 1$
- $Orthogonal(x, y) \cong Sim(x, y) = 0$

We will use these predicates to describe the requirements for similarity functions more clearly. In order to be a similarity measure, the following should hold:

- (reflexive) $Identical(x, x)$

So each object is identical to itself. This is the only requirement for being a similarity measure. Optional properties for similarity measures are:

- (symmetry) $Sim(x, y) = Sim(y, x)$
- (transitive) $Identical(x, y) \wedge Identical(y, z) \Rightarrow Identical(z, z)$
- (indifference) $Orthogonal(x, y) \wedge Identical(z, y) \Rightarrow Orthogonal(x, z)$

Besides similarity, it will also be important to be able to express in what degree two objects are dissimilar. We will use *Dis (x, y)* to denote the dissimilarity between *x* and *y*. In most cases it will be sufficient to define dissimilarity as follows: *Dis (x, y) = 1 – Sim (x, y)* in terms of similarity."

2.3 Cosine Similarity

Documents are often represented as vectors, where each attribute represents the frequency with which a particular term (word) occurs in the document. These are called non-binary vectors. With these vectors, document similarity can be measured. The cosine similarity defined next, is one of the most common measure of documenting similarity. If x and y are two document vectors, then

$$\cos(x, y) = \frac{x \cdot x}{||\mathrm{x}||\ ||y||},$$

where ▪ indicates the vector dot product, x ▪ y = $\sum_{k=1}^{n} x_k y_k$, and ǁxǁ is the length of vector x,

$$\|\mathrm{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x \cdot x}.$$

2.4 Manhattan distance

In the 19th century, Hermann Minkowski considered the so called Taxicab geometry. This is a form of geometry in which the usual distance function of metric or Euclidean geometry is replaced by a new metric. In this new metric, the distance between two points is the sum of the absolute differences of their Cartesian coordinates. [5] The Taxicab geometry has several synonyms. Manhattan distance is one of these synonyms, and it refers to the grid layout of most streets on the island of Manhattan, which causes the shortest path a car could take between two intersections in the borough to have length equal to the intersections' distance in taxicab geometry. The figure bellow, demonstrates different possibilities concerning the paths a car could take to get from point A to point B and vice versa. This approach will be applied in chapter six.

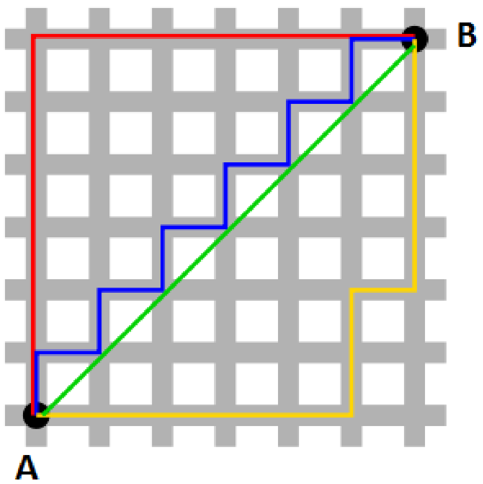

Figure 2.1: Illustration of the Manhattan distance. Red: Manhattan distance. Green: diagonal, straight-line distance. Blue, yellow: equivalent Manhattan distances.
(https://upload.wikimedia.org/wikipedia/commons/thumb/0/08/Manhattan_distance.svg/283px-Manhattan_distance.svg.png)

2.5 Matching by weight - Analytic Hierarchy Process (AHP)

The analytic hierarchy process is a mathematical method/algorithm that can be used to derive ratio scales from paired comparisons. In other words, when combining individual performance indicators to one key performance indicator, each performance indicator can be provided with a weighting. To derive these weightings, the AHP algorithm can be used. This can be helpful when organizing and analysing complex decisions. This can involve decisions that are based on both mathematics and psychology.

AHP was developed by Thomas L. Saaty in 1980. Since then, it has been extensively studied and refined. It has particular application in group decision making, [3]. Throughout the years, AHP has been used in different fields such as business, industry and healthcare for decision making purposes. The algorithm is based on the solution of an eigenvalue problem. In this study, the AHP algorithm is used for the weighting of the features. How this is done will be described elaborately in chapter 4.6.

3. PROBLEM STATEMENT

This chapter provides the necessary background to clarify the relevance of the listed research questions. First, a brief introduction of the main stakeholder 8vance is given, followed by an explicit description of the deep matching technology and its different stages. Then, the deep matching method covered in this study (Strategy Matching), is described in broad lines in section 3.2. In order to clarify the Strategy Matching procedure, an imaging company will be selected which is interested in startup adoption. This is highlighted in section 3.3 together with the accompanying context. This leads to the research questions provided in section 3.4.

3.1 8vance

8vance is a young Dutch software company that was founded in 2013. Its main office is located in Venlo. 8vance has 13 employees at the time of writing.

8vances showpiece is the software system they developed to provide a more efficient alternative to the current way of matching demand (job seekers) and supply (employees) data in the job market. They believe that their system is capable of interpreting data more effective than the technologies e.g. LinkedIn is using. They consider their system as a matching technology, and call it Deep Matching technology[1]. Figure 3.1 gives an illustration of the stages/processes from which this technology exists. In the next paragraphs, a detailed description of these processes is given.

1. Data gathering

The deep matching technology works as follows. Firstly, it needs to be determined what the sources are from which matching data is extracted. Then, unfiltered data is obtained from these sources.

2. Features and sub features

Then, a list of features and related sub features is prepared based on the needs of both the demand and supply parties. This way, relevant information can be filtered from the preciously gathered data. These features can be considered normalized features (will be further referred to as features). In the job market scenario, a few examples of these features would be:

- skills;
- distance (e.g. home to work in km);
- section (e.g. field within which the employee needs to be certified).

Note that two different worlds need to be compared during the matching process. This means that there are two different interpretations of the prepared features. If we use the job market example again, the interpretation of the three features above, would be as follows for the two parties:

Job market example	Interpretation per party	
Features	**Interpretation Supply**	**Interpretation Demand**
Skills	These are the skills I have.	These are the skills we are looking for.
Distance	My job needs to be within x kilometres from my home.	The employee has to reside within x kilometres from working address.
Section	I am certified in field y.	The employee must be qualified in field x or z.

Table 3.1: The job market example.

[1] https://www.8vance.com/en/home/

3. Weighting

After setting up a list of necessary features, the related sub features need to be provided with the corresponding value. This process is called Weighting. If we zoom into the underlying sub features of the example above, then this would result in the following table:

Features	Sub features	Weighting supply party	Weighting demand party
Skills	Accounting	0	1
	Programming	1	6
	Mathematics	1	3
Distance	Five	0	4
	Ten	1	1
Section	IT	1	5
	Finance	0	2

Table 3.2: Relative weighting example

4. Matching

After the weighting process, the so called 'DNA' is ready for matching. For the matching process, different matching algorithms are used. Due to the sensitivity of this information for 8vance, we have decided to exclude the details of the chosen algorithms by 8vance, and the way 8vance implements them. However the algorithms used for the validation of the Strategy Matching approach, are generically disclosed in chapter two. In sections 4.6 and 4.7, a more detailed explanation is given using concrete examples.

Figure 3.1: 8vances DNA deep matching illustration of the job market (https://www.8vance.com/media/imagecontent/process_dashboard_image_2.png)[2]

The numbers used in figure 3.1 are explained in the table below.

Number	Meaning
1	Data gathering
2	Features and sub features
3	Weighting
4	Matching

Table 3.3: Legend deep matching procedure

[2] The illustration used as figure 3.1 is a modified version of the original version which can be found using this URL.

In this research, we will be applying the philosophy explained in this paragraph for the startup adoption market. In figure 3.1, both the supply and demand parties are considered users of the matching platform. In this research, only the corporation searching for startups will be the user of the matching platform. Therefore, this philosophy will be approached differently. In chapter four, a detailed description is given of the matching approach for this project. Then, this approach will be validated in chapter five, followed by a theory proposal of a generic Strategy Matching model in chapter six.

Visualisation

In the current procedure, 8vance transforms the results of this matching procedure into a geographical environment called the smart view. However, the visualisation process is not a part of this research. This part of the procedure will be addressed in a follow-up study. Nevertheless, the visualisation process will be further explained below to give a complete picture of the entire deep matching procedure.

During the visualisation process, the demanding party will be shown in the centre of the smart view with the suppliers around it. Each supplier is been given a colour which indicates the matching percentage. Using this technology ensures that no supplier is excluded. Less interesting suppliers simply get a lower percentage. By avoiding exclusion, you prevent a scenario where less interesting companies are hidden. This way, it is for the demanding party to decide whether a startup with a low matching percentage actually is unsuitable.

To visualize this on a 2D map, the demand is selected as the central point. It is interesting to see how the DM scores (that are expressing a similarity) are being converted to a position in a two dimensional space. The goal is to have a strongly correlated distance between two points on their similarities. Unfortunately, this cannot be solved precisely since there are n attributes which implies to an n dimensional space that needs to be projected on a 2D space using heuristics.

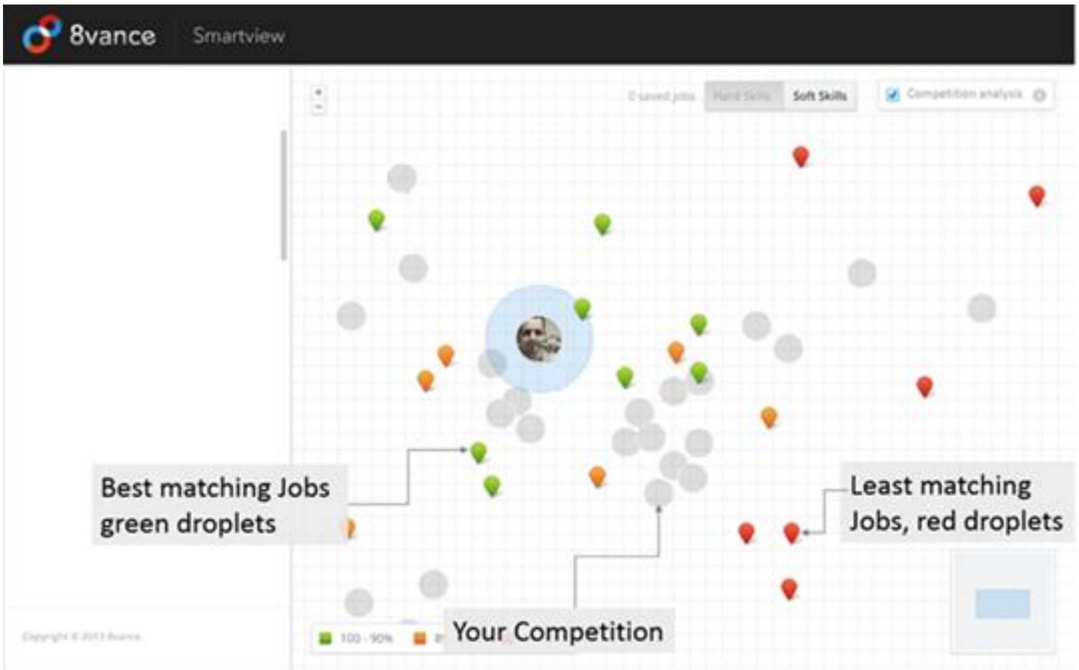

Figure 3.2: Illustration of a smart view (job market) (https://www.8vance.com/media/imagecontent/smart_view_explained_1.png)[3]

The symbols used in the previous illustration are explained in table 3.4.

[3] The illustration used as figure 3.2 is a modified version of the original version which can be found using this URL.

Symbol	Meaning
	70% - 80% match
	80% - 90% match
	90% - 100% match
	Competitors position (not applicable in the Strategy Matching model)

Table 3.4: Smart view legend

3.2 Generic Data Model for Strategy Matching

For the scenario described in section 3.1, 8vance wants to use its deep matching technology to create a generic data model. Consequently, the project is called 'A Generic Data Model for Strategy Matching'. The deep matching technology as described in section 3.1 is universally applicable and can be used for different scenarios. This study shows the possibilities of the deep matching technology for the startup adoption market.

3.3 The Imaging Corporation

In order to be able to validate the deliverables of this study, we will introduce an imaginary company that operates in the imaging market. We will call this company The Imaging Corporation abbreviated as TIC.

Context

TIC is considered a traditional product based imaging company. The past years, TIC has been confronted with a widening of the imaging space and as a result, a rapidly shrinking relevancy. Relatively new trends like servitisation[4] and big data are posing both new opportunities, as well as threats to the more traditional product based companies. For corporate companies like TIC however, there is an opportunity to transform. Weak signals from the startup community could already be matched in an early stage against the corporate technology roadmaps. Their goal is to fine-tune their own strategies or do strategic investments instead of short-term Mergers and Acquisitions (M&A). For TIC, this results in a growing need to find suitable startups for adoption to enable them to make the necessary leaps within their strategic course.

In their search for startups with adoption potential, TIC decided to collaborate with 8vance. 8vance with its Deep Matching Technology™ wants to create an automatic service to help corporations like TIC, understand what is going on in a specific technology startup space and make a connection to relevant startups within that space.

[4] The delivery of a service component as an added value, when providing products. https://en.wiktionary.org/wiki/servitization

3.4 Research questions

Currently, 8vance has different existing models for which they have working matching procedures, including the model for the job market. The goal now is to construct a generic data model (first deliverable) together with its features set (second deliverable). Note that although TIC will act as first launching customer for validating purposes, for 8vance, emphasis is very much on the general applicability of the model. As indicated earlier, this study will result in two deliverables:

- a generic model that serves as the engine of the total procedure;
- a generic features set that serves as the fuel for this engine.

Beneath, the research questions for both the deliverables are listed.

1. To what extent is it possible to establish a model which is suitable for the startup adoption market?
2. What matching procedure is going to be used?
3. What features are required for this model?
4. How are these deliverables going to be validated?

4. FEATURES SET & GENERIC MODEL

In this chapter, the feature set created during this study, will be described elaborately. First, the classification of the features is explained, followed by the explicit description of the features within every class. Then, the weighting process is described for both the matching features as well as their sub features in section 4.6. Lastly, this chapter ends with the detailed explanation of how to determine the final matching result while integrating the weighting process in this calculation.

4.1 Feature classifications

As indicated in the previous chapters, the features have different classifications. Primarily, there are the matching features. These features are used to match the startups with TIC's strategies. Secondly, there are the filtering features. These are features that can rule out particular variables which are currently not relevant to TIC during its search for suitable startups. For example, TIC wants to use the matching algorithm for its EMEA department. Thus, it wants to exclude the other areas. Finally, there are statistical features. These are features that only serve to provide additional information. Information that could help TIC to make a definitive decision concerning a potentially interesting startup. This does not concern data that is being used to match a startup. It only concerns informative data.

What these feature classes, features, and sub features imply in detail, is described elaborately in the chapters that follow. Per class, features and their associated sub features are described in section 4.2 to 4.4.

4.2 Matching features

The features from the matching class have been drawn earlier. In this section, for every feature, an explanation is given together with a description of its adding value and how it should be interpreted.

Startup maturity

This feature reveals how far advanced startups are concerning their maturity. In consultation with the stakeholders, it was decided to let this feature consist of three sub features:

- Validation
- Prototype
- Market

The decision to choose for these three sub features is inspired by NASA's TRL. TRL stands for Technology Readiness Level. This is a method to estimate technology maturity [4]. As shown in figure 4.1, it consists of nine levels. From these nine levels, the three sub features of this feature are distilled.

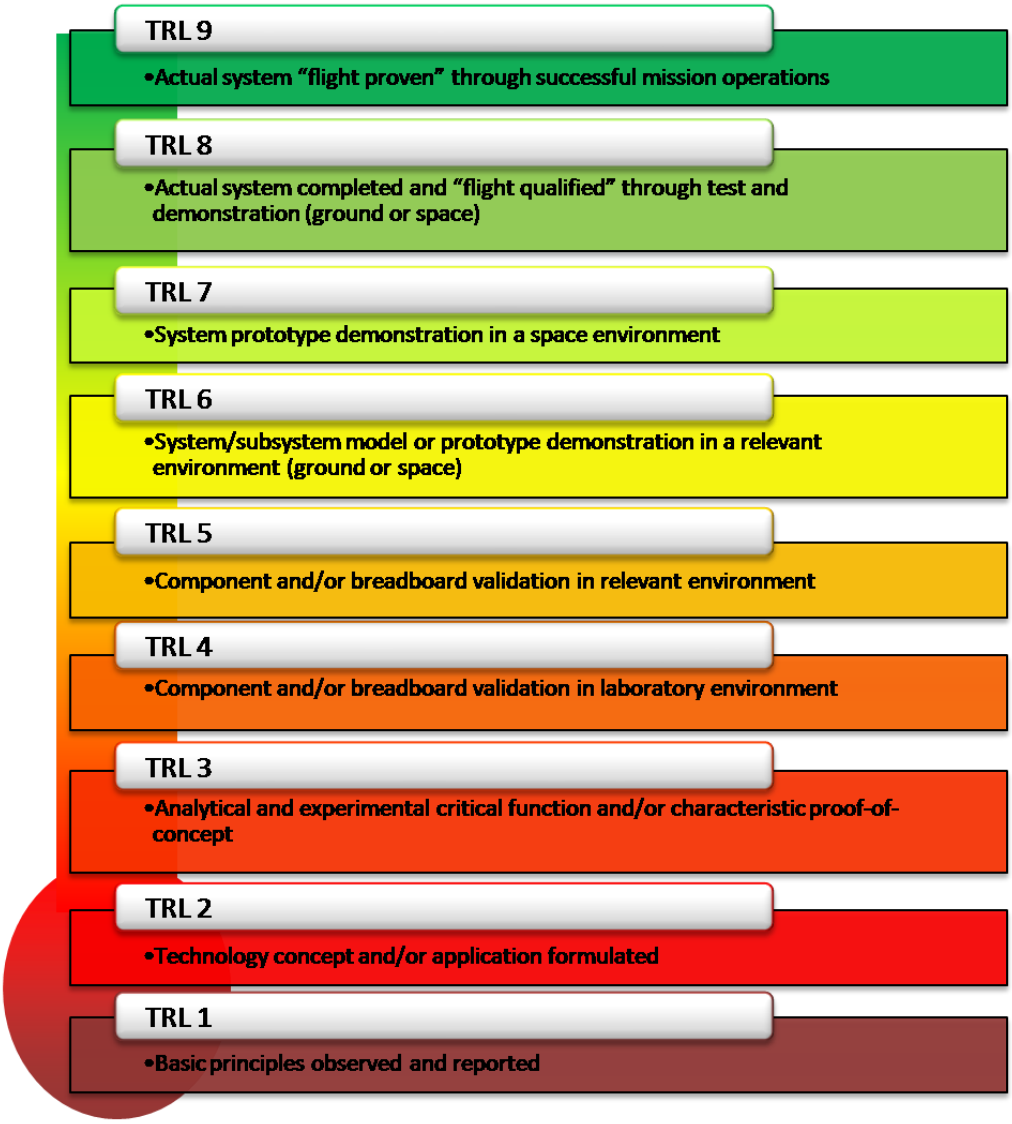

Figure 4.1: Illustration of NASAs Technology Readiness Level.
(https://www.nasa.gov/sites/default/files/trl.png)

The different levels of the TRL are distributed as follows. The Levels 1 to 3 represent Validation, levels 4 to 7 represent Prototype, and levels 8 and 9 represent Market. Figure 4.2 gives a clear overview.

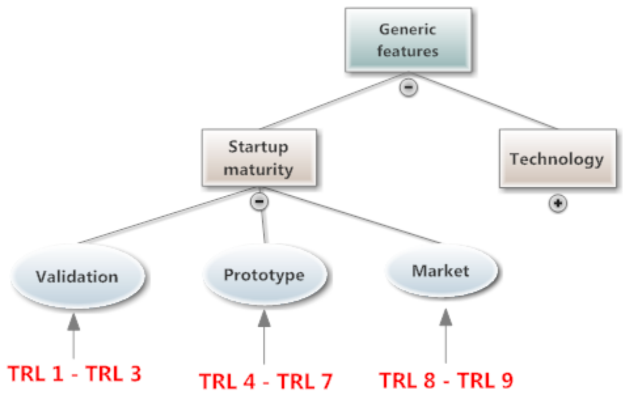

Figure 4.2: Illustration of the Startup maturity feature with the corresponding TRL distribution over the three sub features.

Technology

Technology is one of the most dynamic features. This has two reasons. On the one hand, it is a feature with a constantly growing amount of sub features. On the other hand, it is a feature that can trigger two other features. Namely, the features Technology maturity and Technology maturity speed. This is illustrated in figure 4.3. Like figure 4.2, this is a zoom in from the feature set taxonomy that can be found in Appendix A. The sub feature named X within the Technology feature, is a variable representing every technology that is considered trending.

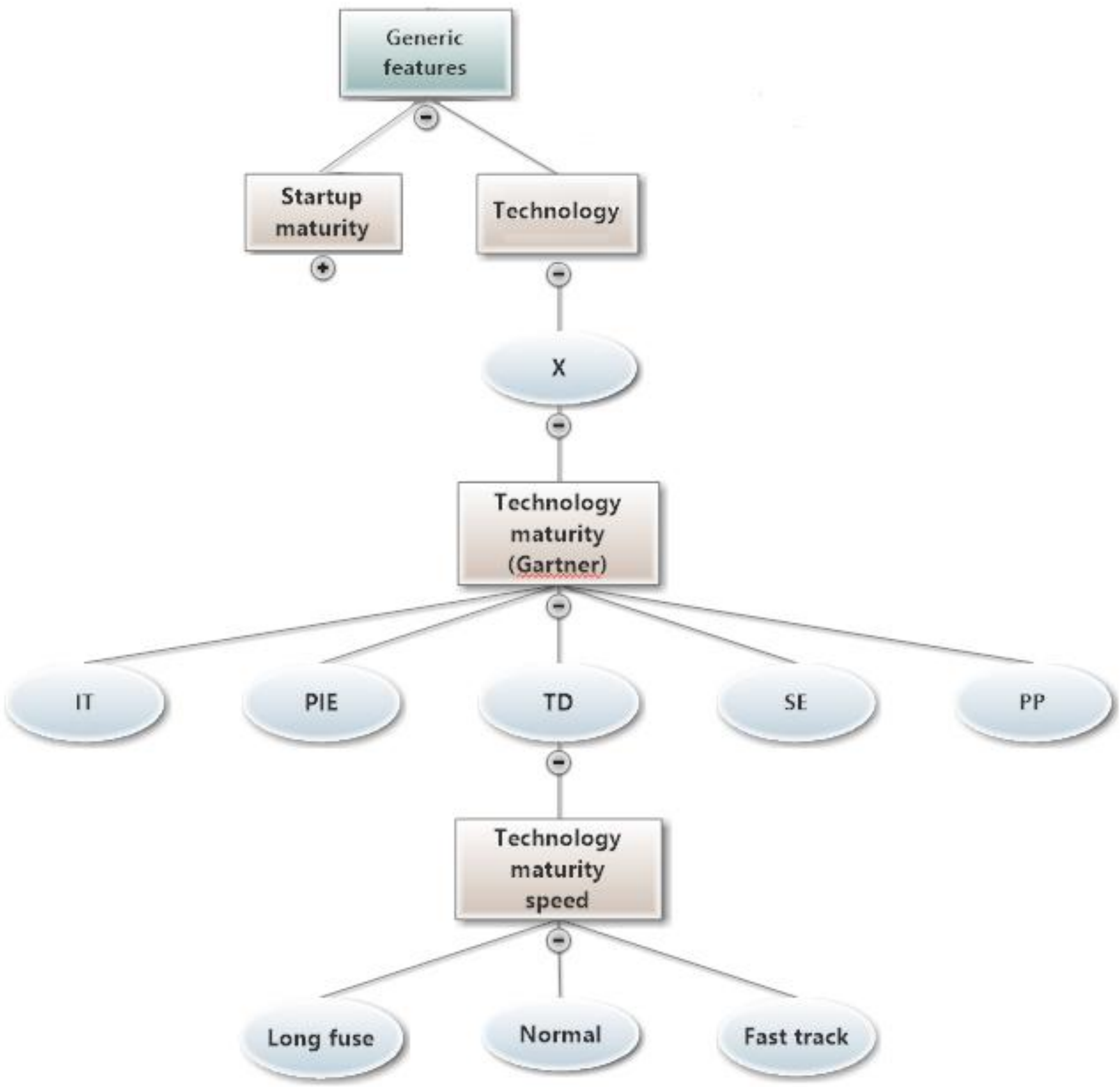

Figure 4.3: Illustration of the Technology feature and the two features that can be triggered when it concerns a trending technology.

At the time of writing, a number of sub features within the Technology feature are considered Trending Technologies. To determine whether a technology is trending or not, the Gartner Hype Cycle is used. This is a report that evaluates the market promotion and perception of value for over 2,000 technologies, services, and trends in over 119 areas. It is a report that tells a story of how technologies, services, and strategies evolve from market hype and excitement of value, to becoming a mainstream part of business and IT [5]. In figure 4.4, an overview is given of the current Trending Technologies.

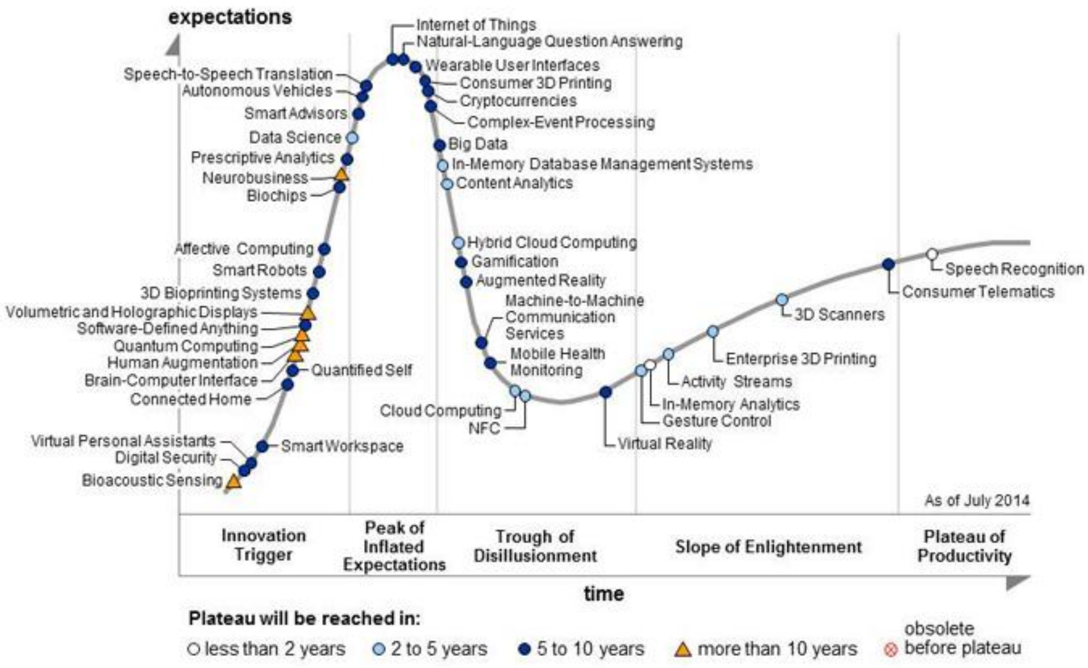

Figure 4.4: Illustration of the Gartner Hype Cycle as of July 2014. (http://na1.www.gartner.com/imagesrv/newsroom/images/HC_ET_2014.jpg;wadf79d1c8397a49a2)

Technology maturity and Technology maturity speed

When a technology is considered trending, the Technology maturity end the Technology maturity speed features are triggered. The Technology maturity feature will indicate in what part of the Gartner Hype Cycle the concerning technology is located, whereas the Technology maturity speed will indicate within which period the concerning technology will pass the Hype Cycle. These are the sub features of the Technology maturity feature:

- Pre IT
- Innovation Trigger
- Peak of Inflated Expectations
- Trough of Disillusionment
- Slope of Enlightenment
- Plateau of Productivity
- Post PP

Pre IT concerns technologies that are recently developed and have not entered the Hype Cycle yet. Post PP covers all the technologies that have passed the Hype Cycle and no longer are considered Trending. These are the sub features if the Technology maturity speed feature:

- Long fuse
- Normal
- Fast track
- None

Long fuse concerns technologies that stay in the Hype Cycle for one or two decades. Normal is the speed indicator for technologies that take five to eight years until they pass the Cycle, and Fast track concerns technologies that leave the Hype Cycle within two to four years. When a technology is not considered trending, it will be describes as Pre IT or Post PP in the Technology maturity feature. When this is the case, it will get the sub feature 'None' in the Technology maturity speed feature.

Business type

Business type helps to determine what type of service a startup provides. The sub features this feature consists of are:

- Hardware
- Software
- Services
- None

Hardware implicates that the startup is selling physical products. Software concerns selling virtual products. Services applies for organisations that are willing to go a bit further than what is considered standard (e.g. a 24/7 helpdesk). The sub feature None means the startup does not focus on any of the other sub features mentioned in this paragraph.

Business target

The Business target feature helps TIC to determine which customer segment a startup serves. The sub features that are present in the Business target feature are as follows:

- Mass market
- Niche market
- Segmented market
- Diversify
- Multisided

Mass market covers the general users (e.g. MS Office Word) whereas Niche market focusses on specific users (e.g. LaTeX). Segmented market can be considered as a distinction of niche. Diversify focusses on several costumer segments with different needs (e.g. Google or Apple). Lastly, a multisided company can be described as an intermediate company that connects multiple niches (e.g. Uber or Ebay).

Business target relation

This feature consists of four sub features:

- Business to Business (B2B)
- Business to Consumer (B2C)
- Consumer to Business (C2B)
- Consumer to Consumer (C2C)

Business target relation helps to clarify which transactions startups facilitate through their offer. Since the names of these sub features speak for themselves, only companies will be added that may serve as an example of every sub feature. B2B facilitating company examples would be Shutterstock or EyeEm. However, EyeEm is also a good example of B2C. Another good example of B2C would be Canary. Moreover, EyeEm is also well fit as a C2B facilitating company. Another example of a C2B company is iStockphoto. Lastly, Ebay and Marktplaats are good examples of companies that facilitate C2C environments.

Revenue
Revenue's main purpose is to determine what source(s) of income startups have. Since there are more than five sub features present in Revenue, table 4.1 is added to ensure a proper overview.

Feature	Sub features	Description/example
Revenue	Asset sale	Hardware purchase
	Usage fee	Per x
	Subscription	Software per timeframe
	Lending	Hardware per timeframe
	Licensing	Software purchase
	Brokerages fee	Intermediate costs (e.g. Uber)
	Advertisement	Free, Advertisement fee
	Free	Free of not public

Table 4.1: Revenue feature table with the relating sub features and their descriptions/examples.

4.3 Filtering features

At the time of writing, the following three filtering features are in use:

- Area
- Startup status
- 2005 or younger

In the next paragraphs, a brief description of these features and their sub features is provided.

Area
The Area feature indicates in which area a startup is located. Here, the focus is on the headquarters. The sub features associated with this feature are:

- Japan
- America
- Asia
- EMEA

It goes without saying that Japan is located in Asia but for strategic reasons, TIC decided to include Japan separately from her continent. Furthermore, when referring to America, TIC means North, Central, and South America. EMEA stands for **E**urope, **M**iddle **E**ast, and **A**frica. Recently, TIC also has added Australia to this area.

For this project the focus is EMEA. Nevertheless, information on startups in other areas is also considered relevant for analytical purposes. When on a particular moment TIC has no interest in other areas than EMEA, these areas can simply be excluded using this filtering feature.

Startup status
This feature describes what the current market position is of a startup. The sub features defining the different positions are:

- Acquired
- Operating
- Closed
- IPO

Acquired indicates that the corresponding startup is adopted by another organization. Operating means that the startup is still active and has not been acquired by a third party. Closed indicates that the startup is no longer active and is therefore of no interest for acquisition purposes. IPO indicates that the startup has been listed. This makes the startup often less attractive for adoption since the market value of the startup rises significantly.

10 years or younger
For all the organizations that want to apply the matching algorithm, it is important to clarify what their definition of a startup is regarding to the temporal length of its existence. TIC's interpretation of this is a company that is ten years old or younger. Hence, the feature '10 years or younger'. The sub features of this feature are simply Yes and No.

4.4 Statistical features
This feature set currently exists of three features:

- Crunchbase
- Kickstarter
- IndiGoGo

Crunchbase
The information displayed in the various sub features of this feature is derived from the Crunchbase.com website. It involves the following sub features:

- Funding round type
- Funding round code (Series)
- $ raised
- Valuation type
- Announced on
- Closed on
- Investors
- Date checked

Crowd funding
The information corresponding to the sub features below comes from the crowd funding websites kickstarter.com and indigogo.com. This information will be displayed only when a startup has started a crowd funding project on one or both of these websites. The sub features associated with the Crowd funding feature are:

- # Funders
- Flexible funding? (IndiGoGo)
- $ Goal
- $ Raised
- % Raised
- Start date
- End date
- Date checked
- Success?

On the basis of this information, TIC can deduce how popular the concerning product is with consumers (sentiment data).

Investment fit
This feature gives additional information about the startup. The founder of a startup is the main focus in this feature. The sub features in the Investment fit feature are based on data that TIC considers being of potential value. These are the concerning sub features:

- Right person
- Right idea
- Right product
- Right time
- Right market

By 'Right person', the founder of the concerning startup is meant. In TIC's interpretation, the relevant data for this sub feature is the track record of the founder en where he is graduated. Whether it is the right time for this startup depends on the technologies this startup is using, and whether these are trending technologies (Gartner Hype Cycle). Whether it is the right market can be concluded based on the amount of money raised by the startup.

4.5 Recap features set
The features set consists of three classes. As the name indicates it, the features within the matching features class are used for matching purposes. Filtering features can be used to exclude startups based on e.g. area or startup status. Features within the statistical features class serve only to provide additional information. This concerns information that could help the startup adopting corporation to make a definitive decision concerning a potentially interesting startup. Note that this is purely informative data, and it does not concern data that is being used to match a startup.

4.6 Weighting the matching features and sub features
In this section, the process of giving a hierarchical structure to the matching features is explained. The explanation is done by taking three nonexistent features. In the last paragraph of this section, the weighting process of the sub features is described.

Adding weights to features

First of all, the matching features need to be provided with a weight in an x/x comparison matrix. In order to determine the proper relative weighting for all the matching features, the features are being weighted by comparing them pairwise. The example in table 4.2 illustrates this procedure in a three by three comparison matrix M:

M	a	b	c
a	1	2	5
b	$\frac{1}{2}$	1	2
c	$\frac{1}{5}$	$\frac{1}{2}$	1

M[a,b] = 2 means a is two times more important than b and M[c,b] = $\frac{1}{2}$ means c is 0.5 times more important than b etc.

Table 4.2: Example to illustrate the relative weighting process using a comparison matrix.

We use the scaling scheme conceived by Saaty and adapted by Paraskevopoulos's [7] as shown in table 4.3. A consistency requirement is that when e.g. feature a is 2 times more important than b, this automatically means b is $\frac{1}{2}$ more important than a etc.

Intensity of importance	Definition	Explanation
1	Equal importance	Two factors contribute equally to the objective.
3	Somewhat more important	Experience and judgment slightly favour one over the other.
5	Much more important	Experience and judgment strongly favour one over the other.
7	Very much more important	Experience and judgment very strongly favour one over the other. Its importance is demonstrated in practice.
9	Absolutely more important	The evidence favouring one over the other is of the highest possible validity.
2, 4, 6, 8	Intermediate values	When compromise is needed.

Table 4.3: Table to explain the scaling used for the relative weighting of the features.

Eigenvector

Our goal is to determine the absolute weighting of the matching features from the comparison matrix. In AHP, this is done by deriving the principal eigenvector from the relative weighting (comparison matrix). There are numerous approaches for deriving the eigenvector. Paraskevopoulos's [7] indicates, "Multiplying together the entries in each row of the matrix and then taking the n^{th} root of that product gives a very good approximation to the correct answer. The n^{th} roots are summed and that sum is used to normalize the eigenvector elements to add to 1.00." In table 4.4, we illustrate this by extending the example of table 4.2.

M	a	b	c	n^{th} root of products of values	Eigenvector
a	1	2	5	2.154	0.595
b	$\frac{1}{2}$	1	2	1.000	0.276
c	$\frac{1}{5}$	$\frac{1}{2}$	1	0.464	0.128
SUM				3.619	1.000

Table 4.4: Extension of the comparison matrix example illustrating the derivation of the eigenvector.

Transitivity

If the comparison matrix used in table 4.2 would be transitive, this would mean that $M[a,b] \cdot M[b,c] = M[a,c]$ for all attributes a,b,c. However, this comparison matrix is not transitive. We have done this to show that the weighting does not need to be 100% transitive in order to be able to compute the attribute weights. Note that when a comparison matrix is not transitive, the consistency ratio (CR) must remain below 10% since Saaty [8] indicates that a CR above 10% may be too inconsistent to consider reliable. To determine the CR, we need to divide the consistency index (CI) with the corresponding index of consistency for random judgments (RI). The RI can be derived from Saaty's book [8]. The CI however, can be calculated using the following formula:

$$CI = (\lambda_{max} - n)/(n-1)$$

where n are the elements to be compared (features). Thus, in order to lead to the CI and CR, we need to calculate the principal eigenvalue λ_{max} first. In [7], a detailed explanation is given of how to determine the λ_{max}.

Adding weights to sub features

In order to calculate the absolute weighting, the relative weighting of the sub features needs to be determined. This is done based on a selected strategy. For the relative weighting of the sub features, we decided to divide one point over all the sub features of a feature. Table 4.5 is an example that illustrates this procedure. For this example, the Business type feature is selected, and the relative weighting quantification is fictitious.

Feature	Business type			
Sub features	Hardware	Software	Services	None
Relative weighting	0.2	0.3	0.5	0

Table 4.5: Example, Business type feature - relative weighting of its sub features.

In addition to the weighing of the sub features, also the startups need to be scouted. This means that every startup will get the sub feature values associated with that particular startup. This scouting process is completely objective, since it is based on the properties of the startups. The example in table 4.5 could be completed as followed:

Feature	Business type			
Sub features	Hardware	Software	Services	None
Relative weighting	0.2	0.3	0.5	0
Startup X	0	0.1	0.9	0

Table 4.6: Example, Business type feature - relative weighting of its sub features plus the objective values of Startup 'X' for this sub feature.

4.7 Matching results

In order to determine the matching result between a selected strategy and a startup from the features as described above, still a number of steps need to be followed. First, the absolute weighting of the features needs to be merged with the relative weighting of the sub features by multiplying every feature vector component with its relating sub features. This has to be done both for the selected relative weighting and the startup in question separately. If we use the vector component of feature 'a' from table 4.4 to illustrate this procedure, this would result in the following calculation:

Feature	a (Business type)			
Sub features	Hardware	Software	Services	None
Relative weighting	$0.595 * 0.2$	$0.595 * 0.3$	$0.595 * 0.5$	$0.595 * 0$
Startup X	$0.595 * 0$	$0.595 * 0.1$	$0.595 * 0.9$	$0.595 * 0$

Table 4.7: Eigenvector multiplication with relative weighting of sub features and startups.

This way, the features vector is included in the final matching process. Together with the matching process of the sub features, it will help determine the final matching result. After the multiplication, the matching process of the sub features can start. Before calculating the final matching result, we need to normalize the results from table 4.7 first.

Feature	a (Business type)			
Sub features	Hardware	Software	Services	None
Relative weighting	$\frac{0.119}{\sqrt{0.119^2+0.179^2+0.298^2+0^2}}$	$\frac{0.179}{\sqrt{0.119^2+0.179^2+0.298^2+0^2}}$	$\frac{0.298}{\sqrt{0.119^2+0.179^2+0.298^2+0^2}}$	$\frac{0}{\sqrt{0.119^2+0.179^2+0.298^2+0^2}}$
Startup X	$\frac{0}{\sqrt{0^2+0.060^2+0.540^2+0^2}}$	$\frac{0.060}{\sqrt{0^2+0.060^2+0.540^2+0^2}}$	$\frac{0.540}{\sqrt{0^2+0.060^2+0.540^2+0^2}}$	$\frac{0}{\sqrt{0^2+0.060^2+0.540^2+0^2}}$

Table 4.8: Normalized values

Next, the improduct can be determined. The improduct can be considered as the final matching result and can be calculated as follows:

$$Inproduct = 0.324 * 0 + 0.487 * 0.110 + 0.811 * 0.994 + 0 * 0$$

Finely, to express the extent to which a startup matches with the weighted strategy in percentages, the matching results can be multiplied by 100%. For the feature example we have used so far, this would be:

$$Inproduct \ * \ 100\% \ = \ 0.860 \ * \ 100\% \ = \ 86\%$$

5. VALIDATION

After the matching approach for both the weighting elements were chosen, it is necessary to test the functionality of the model in practice. By functionality, the application of the model is meant in 8vance context. How this process is handled, will be described in this chapter step by step.

Potential startups

For a proper testing environment, it is necessary to have a number of startups that can be screened. Thus, the TIC stakeholder has been asked to set up a list of startups that are operating within the imaging market. This resulted in a selection of 11 startups with different strategic angles within the imaging market. These are the selected startups listed alphabetically:

- Canary
- Evercam
- EyeEm
- Frontback
- Meerkat
- Narrative
- Orbeus
- Podo
- Storehouse
- Triggertrap
- Withings

Ranking startups

To validate the Strategy Matching approach introduced in this study, we have decided to rank the listed startups based on the extent to which they match with a TIC strategy, and to compare this with an intuitive ranking provided by the TIC stakeholder. Next to these two rankings, we have also created a completely fictitious ranking based on the '=RAND()' function provided in MS Excel. Needless to say, we used exactly the same startup values as used for the other rankings. As explained in section 4.6, these values are obtained during an extensive scouting of these startups. Table 5.1 shows this fictitious ranking.

Startups	Ranking	Matching %
Evercam	1	98,32%
Triggertrap	2	97,37%
Narrative	3	86,95%
Meerkat	4	85,75%
Withings	5	73,77%
EyeEm	6	64,37%
Podo	7	50,53%
Canary	8	49,10%
Storehouse	9	37,14%
Orbeus	10	20,67%
Frontback	11	19,56%

Table 5.1: Random ranking of the selected startups.

5.1 TIC strategy – intuitive ranking

Like every corporation, TIC has different strategies. This means that the matching feature set and the underlying sub features may have different hierarchical structures. Bear in mind that this has no effect on the algorithms used. It only results in different weightings which consequently will affect the matching results.

Therefore, the TIC stakeholder was asked to select a TIC strategy and to make an intuitive ranking of the startups based on the selected strategy. Also, the stakeholder is asked to give matching percentages to make his estimations more tangible. These percentages will not be considered as a strict measurement tool during the comparison of the intuitive ranking and the Strategy Matching ranking, since the intuitive percentages are only relative. They will only contribute in obtaining a general idea of the intuitive ranking. The strategy selected for the matching procedure of this validation phase is called Imaging Universal. A concise description of the selected strategy will follow in the next paragraph.

Imaging Universal strategy

A general description of imaging is that it can be seen as the representation of reproduction of an object's form; especially a visual representation (i.e., the formation of an image)[5]. When we look at imaging in the context of TIC as a versatile imaging corporation, it is reasonable to assume that there are various directions within this business. This has divided this into two main categories: Business Imaging and Consumer Imaging. For the validation, TIC selected a strategy that focusses on both categories. Hence, the title, Imaging Universal.

Startups	Ranking	Matching %
EyeEm	1	70,00%
Withings	1	70,00%
Narrative	3	65,00%
Triggertrap	3	65,00%
Canary	5	60,00%
Orbeus	5	60,00%
Evercam	7	55,00%
Frontback	7	55,00%
Storehouse	7	55,00%
Meerkat	10	50,00%
Podo	11	35,00%

Table 5.2: Intuitive ranking - Imaging Universal

5.2 Strategy Matching

Next to a random ranking and an intuitive ranking, we also need to rank these startups using the Strategy Matching approach developed during this study. To complete the validation procedure, these rankings will be compared, and a conclusion will be drawn from the results. How the ranking

[5] https://en.wikipedia.org/wiki/Imaging

of the startups based on the Strategy Matching approach will shape, is explained in the two following paragraphs.

Relative weighting - Imaging Universal

The TIC stakeholder was requested to give the proper relative weighting to the features based on the selected strategy using the AHP approach as explained in section 4.6. Table 5.3 shows the comparison matrix of the current features containing the relative weighting based on the Imaging Universal strategy.

Features	Startups maturity	Technology	Technology maturity	Technology maturity speed	Business type	Business target	Business target relation	Revenue
Startups maturity	1	1/7	1/3	1/5	1/2	1/3	1/2	2
Technology	7	1	2 1/3	1 2/5	3 1/2	2 1/3	3 1/2	14
Technology maturity	3	3/7	1	3/5	1 1/2	1	1 1/2	6
Technology maturity speed	5	5/7	1 2/3	1	2 1/2	1 2/3	2 1/2	10
Business type	2	2/7	2/3	2/5	1	2/3	1	4
Business target	3	3/7	1	3/5	1 1/2	1	1 1/2	6
Business target relation	2	2/7	2/3	2/5	1	2/3	1	4
Revenue	1/2	1/14	1/6	1/10	1/4	1/6	1/4	1

Table 5.3: Relative weighting of the features based on a chosen TIC strategy.

Ranking feature vector components

After the derivation of the relative weighting, we can start determining the absolute weighting. This means, determining the feature vector components using the AHP algorithm as described in chapter 4.6. Consequently, the resulting vector components are as listed in table 5.4. By ranking these vector components, we can show the hierarchical order of the relating features.

Features	Vector components	Ranking
Startups maturity	0,043	7
Technology	0,298	1
Technology maturity	0,128	3
Technology maturity speed	0,213	2
Business type	0,085	5
Business target	0,128	3
Business target relation	0,085	5
Revenue	0,021	8

Table 5.4: Results of the absolute weighting based on the selected strategy.

The TIC stakeholder was also requested to translate the chosen strategy (Imaging Universal) to the relative weighting composition of the sub features. The way this needs to be done, is described in the paragraph 'Adding weights to sub features' of section 4.6.

These relative weightings then were used to match every startup using the concerning algorithms as described in sections 4.6 and 4.7. Lastly, the matching results were listed in a ranking table similar to the intuitive ranking table (table 5.2). For the Imaging Universal strategy, the matching results were as follows:

Startups	Ranking	Matching %
Orbeus	1	85,23%
Narrative	2	83,75%
Triggertrap	3	83,64%
Withings	4	83,23%
Podo	5	78,63%
Canary	6	76,93%
EyeEm	7	75,06%
Evercam	8	73,34%
Frontback	9	26,87%
Storehouse	10	26,44%
Meerkat	11	20,21%

Table 5.5: Strategy Matching ranking - Imaging Universal

5.3 Comparing the rankings

When comparing the intuitive ranking with the results of the Strategy Matching approach, there are a number of outcomes to be noticed. This concerns both similarities as well as contradictions. Next to these tables, also the random ranking has been provided to accentuate the adding value of the Strategy Matching approach.

Contradictions

The first thing we noticed, is the position of EyeEm, Orbeus and Podo in the intuitive ranking compared to their position in the Strategy Matching ranking. We believe that this could indicate that these startups may not have been provided with the proper sub feature values. This can be analysed with the TIC stakeholder. It may also indicate that the TIC stakeholder overestimates the added value of EyeEm and/or underestimates the added value of Orbeus and Podo for the Imaging Universal strategy. A major contradiction between the random ranking and the other two rankings, is the almost 100% match of Evercam in the random ranking.

Startups	Ranking	Matching %
EyeEm	1	70,00%
Withings	1	70,00%
Narrative	3	65,00%
Triggertrap	3	65,00%
Canary	5	60,00%
Orbeus	5	60,00%
Evercam	7	55,00%
Frontback	7	55,00%
Storehouse	7	55,00%
Meerkat	10	50,00%
Podo	11	35,00%

Startups	Ranking	Matching %
Orbeus	1	85,23%
Narrative	2	83,75%
Triggertrap	3	83,64%
Withings	4	83,23%
Podo	5	78,63%
Canary	6	76,93%
EyeEm	7	75,06%
Evercam	8	73,34%
Frontback	9	26,87%
Storehouse	10	26,44%
Meerkat	11	20,21%

Startups	Ranking	Matching %
Evercam	**1**	**98,32%**
Triggertrap	2	97,37%
Narrative	3	86,95%
Meerkat	4	85,75%
Withings	5	73,77%
EyeEm	6	64,37%
Podo	7	50,53%
Canary	8	49,10%
Storehouse	9	37,14%
Orbeus	10	20,67%
Frontback	11	19,56%

Table 5.6: Contradictions - the intuitive, Strategy Matching and random ranking are shown respectively.

Similarities

When looking to the intuitive and the Strategy Matching ranking, the rest of the order in which the other startups follow is very similar. For a better overview, we have decided to separate the startups that show similarity in the upper part of the ranking, from the startups in the lower part, by highlighting them with different colours. In the upper part of the ranking, there is a minor difference to be noticed concerning the similarity. This may indicate that further research is necessary to fine tune the weighting in order to improve its quality. Therefore, this approach is included in the 'Further research' section of chapter seven. However, it could also indicate that the intuitive ranking is less effective e.g. due to a confounding factor. This will be covered in the next paragraph. Lastly, in the random ranking, the two highlighted groups of startups, seem to be mixed and no logical similarity can be found whatsoever.

Startups	Ranking	Matching %	Startups	Ranking	Matching %	Startups	Ranking	Matching %
EyeEm	1	70,00%	Orbeus	1	85,23%	Evercam	1	98,32%
Withings	1	70,00%	Narrative	2	83,75%	Triggertrap	2	97,37%
Narrative	3	65,00%	Triggertrap	3	83,64%	Narrative	3	86,95%
Triggertrap	3	65,00%	Withings	4	83,23%	Meerkat	4	85,75%
Canary	5	60,00%	Podo	5	78,63%	Withings	5	73,77%
Orbeus	5	60,00%	Canary	6	76,93%	EyeEm	6	64,37%
Evercam	7	55,00%	EyeEm	7	75,06%	Podo	7	50,53%
Frontback	7	55,00%	Evercam	8	73,34%	Canary	8	49,10%
Storehouse	7	55,00%	Frontback	9	26,87%	Storehouse	9	37,14%
Meerkat	10	50,00%	Storehouse	10	26,44%	Orbeus	10	20,67%
Podo	11	35,00%	Meerkat	11	20,21%	Frontback	11	19,56%

Table 5.7: Similarities – the intuitive, Strategy Matching and random ranking are shown respectively.

Confounding factors

Next to the validation of the weighting approach and the algorithms that help to determine the matching result, some parts of the Strategy Matching simply cannot be validated. It is important to be transparent about this fact so this can be taken into account during quality evaluation.

As explained earlier, during the relative weighting phase, the TIC stakeholder is asked to provide both the features and the sub features with relative weighting based on the selected strategies. Bear in mind that this process is intuitive and therefore sensitive to various external variables like e.g. emotional state, level of focus, and workload of the individual translating these strategies into relative weighting. This also applies for the startup scouting procedure. During this procedure, the quality of the startups need to be transformed into sub feature values. This also is an intuitive technique making it sensitive to the variables mentioned earlier.

6. G8MT – THEORY PROPOSAL

In this chapter, the generic data model is described. In order to do this step by step, the Seligman (Seligman, Wijers, & Sol, 1989) framework will be used. This framework consists of the components shown in the info graphic below.

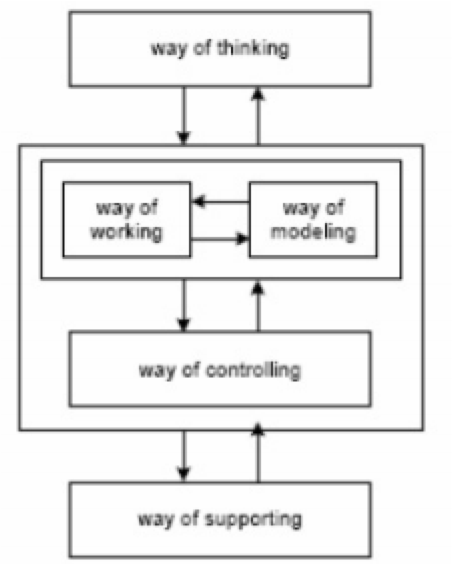

Figure 6.1: Framework to describe methods explicitly (Seligman et al., 1989)

For this study, 'A way of working' is equivalent to chapter four. Therefore, this will not be further discussed in this chapter. In the following sections, we will restrict ourselves to the remaining relevant parts of this framework. Namely, 'A way of thinking' and 'A way of modelling'.

6.1 General perspective

Following a comprehensive description of the current situation in chapter three followed by the elaborate description of the deliverables, it is now important to focus on the applied modelling technique. Prior to this focus, we need to clarify what our perspective is. This step can be considered as the way of thinking.

Complex objects are described by their attributes. A description framework describes what attributes are used and what the attribute values can be. Attributes may relate to complex sub-objects, described by an associated description framework or may have an elementary value.

6.2 Objects and their matching potential

In this section, the chosen approach will be formalised considering the generic model described in the previous section. Here, the essential concepts and their relations are elucidated. Therefore, this step can be seen as the way of modelling.

The generic 8vance modelling technique (G8MT)

We assume a context where various similar objects are to be compared, and the similarity between the objects is used to cluster the objects. Using standard techniques, such clustering is effectively displayed and can be used as a decision support tool.

The generic 8vance modelling technique (G8MT) is used to make a description framework for the objects under consideration. This framework is a feature hierarchy that allows for an effective grouping and classification of features on various levels of granularization. Setting up the description framework is the major challenge when modelling some application domain.

A description framework consists of a set of attribute-value pairs. A value may be either elementary or a description framework. We describe this structure by the following context free grammar:

DF → *Name* = *PRL*	*//name + list of attribute value pair*
PRL → *PR* | *PR, PRL*	*//construction of the list*
PR → *A:D* | *A:DF*	*//attribute - value pair option*
D → *Standard Domain*	*//underlying atomic domain*

where the syntactic category *DF* corresponds to the description framework, *PRL* to a property list, *PR* to an individual property, *A* to an attribute name and *V* to the associated value domain. An example is the following simple scheme:

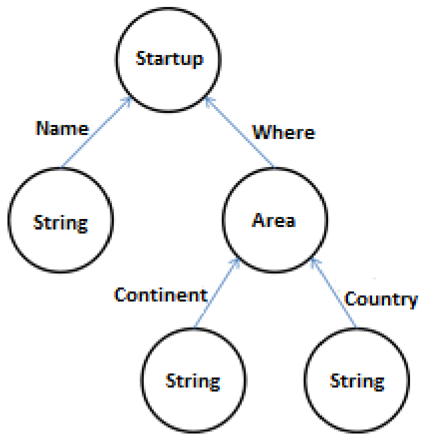

Figure 6.2: Startup scheme

which is formally described as:

Startup = Name: String, Where: Area = Continent: String, Country: String

Let *x* be an instance of this description framework, then it follows that *x* has the following structure:

1. *Def (x)* is the set {*Name, Where*}
2. *x (Name)* is a string value
3. *x (Where)* is an instance of the part of the description framework headed by *Area*
4. *Def (x (Area))* is the set {*Continent, Country*}
5. *x (Where) (Continent)* is a string value
6. *x (Where) (Country)* is a string value

Deep Matching

Deep matching computes the similarity between two objects, taking the structure of the description framework into account. We will now describe how similarity is determined using deep matching.

$$Sim_{\ Manhattan}(x, y) = 1 - \frac{|Def(x) \div Def(y)|}{|Def(x) \cup Def(y)|}$$

Given the objects x and y illustrated in figure 6.3, there is a difference in definition between the domains Def(x) and Def(y) respectively of the two objects. To determine the similarity between these objects based on this domain difference, we can use the Manhattan similarity approach:

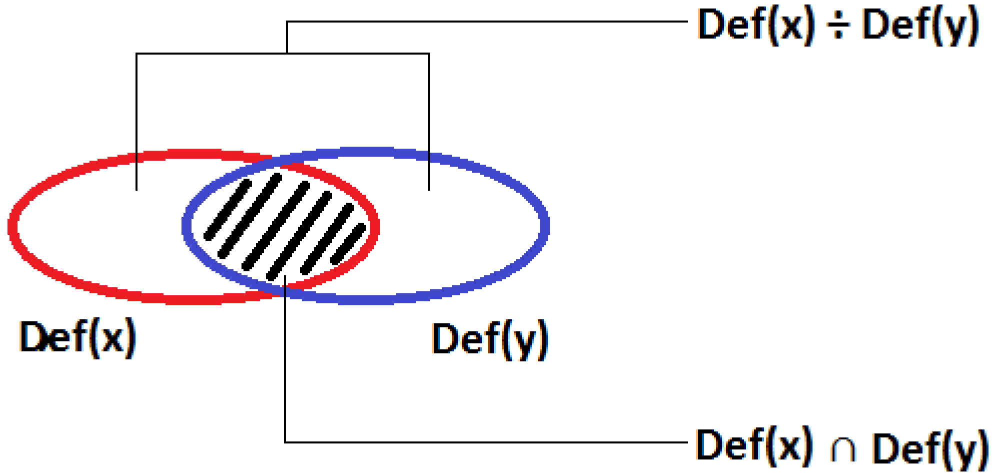

Figure 6.3: Example similarity between two objects.

Then, what remains are the attributes for which both x and y have associated in value. In figure 6.3, this would be $Def(x) \cap Def(y)$ which from now on will be written as $C(x, y)$. To determine the similarity based on this part, the following generic formula can be used:

$$Sim\ (x, y) = \frac{commonality\ between\ x\ and\ y}{generality\ of\ x\ and\ y}$$

For this purpose, we will define commonality as follows:

$$Comm_{G8MT}\ (x, y) = \sum_{f \in C(x,y)} w_f \cdot \begin{cases} Sim_{Dom(f)}\ (x(f), y(f)) & \text{if } Def\left(x(f)\right) = \emptyset \\ Comm\ \ (x(f), y(f)) & \text{otherwise} \end{cases}$$

For the generality, we use $\sqrt{|x| \cdot |y|}$ (which is also used in the Cosine measure). The w_f is a weight factor for features. In section 4.6, we have described how such weights may be determined in practice easily by using the AHP method.

Now, both parts ($Sim_{Manhattan}$ and $Comm_{G8MT}$) can be combined using a weighting factor α. Choosing the best value for α, can be done based on prior knowledge or by trial and error.

$$Sim\ (x, y) = \alpha \cdot \frac{Comm_{G8MT}\ (x, y)}{\sqrt{|x| \cdot |y|}} + (1 - \alpha) \cdot Sim_{Manhattan}(x, y)$$

Clustering objects

As defined in [8], the purpose of clustering objects is to determine the intrinsic grouping in a set of unlabelled data. There are two types of clustering, supervised and unsupervised. [9] Eick et al. indicate "Clustering is typically applied in an unsupervised learning framework using particular error functions, e.g. an error function that minimizes the distances inside a cluster keeping clusters tight. Supervised clustering, on the other hand, deviates from traditional clustering in that it is applied on classified examples with the objective of identifying clusters that have high probability density with respect to a single class."

In this study, we have chosen for an unsupervised training in which network learns to form their own classifications of data without human input. A most popular technique for this training method is Teuvo Kohonen's Self-Organising Maps, abbreviated as SOM. The technique supposes a distance measure d(x, y) to indicate the (conceptual) distance between objects x and y. In the next section we will discuss the relation between this distance measure and the similarity measure from the previous section.

As explained above, SOM is the technique selected for clustering in this study. In [10], SOM is described as a neural network algorithm whose main goal is to transform an incoming signal pattern of arbitrary dimension into a one or two dimensional discrete map. Moreover, this transformation needs to be performed adaptively in a topologically ordered fashion. The operation of SOM can be described as follows:

```
wj := random initial weight
repeat

    x := sample training from input space          //sampling
    I := neuron with weight vector closest to x    //matching
    Δ wj := η(t) Tj,I (t) (xi − wj)                //updating

until feature map stable
```

Similarity versus distance

We now have similarity. The final step is to determine the distance from the similarity. First, a brief introduction is given to describe distance within this context.

In [2], "Distance functions are very different from similarity functions. The distance *d (x, y)* between vectors *x* and *y* is expressed as a positive real number. The following properties are required for distance functions:

1. (lower bound) $d(x, x) = 0$
2. (different vectors have a distance) $x \neq y \Rightarrow d(x, y) > 0$
3. (symmetry) $d(x, y) = d(y, x)$
4. (triangular inequality) $d(x, y) + d(y, z) \geq d(x, z)$"

Now, we will explain how we relate similarity and conceptual distance. As [2] indicates, "Similarity and distance are very different ways to relate objects to each other. Their relation should satisfy the following requirements:

1. A smaller distance corresponds to a larger similarity
2. A larger distance corresponds to a smaller similarity.

Furthermore, small variations in distance have more impact on similarity for small distances than for large distances. For example, an age difference of 1 is more significant for children than for elderly people.

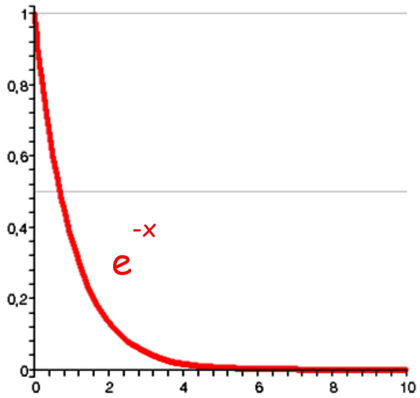

Figure 6.4: Possible relation between distance and similarity.

As a consequence, the relation between similarity and distance may be defined as follows:

$$Sim\ (x, y) = e^{-d\ (x,y)}$$

If *d* is a distance measure, then it is easily verified that *Sim* is a similarity measure. So if we have given a similarity function *Sim*, then the associated distance function *d* in this approach would be:

$$d(x, y) = -\ln(Sim(x, y))$$"

7. CONCLUSION & FURTHER RESEARCH

In this chapter, both the conclusion of this study and advice regarding follow-up studies are included respectively followed by a summarization of the contributions of this research.

7.1 Conclusion

More and more companies choose to adopt other, typically startup organizations, with the ambition to make a leap forward in their strategic direction. This study provides a generic model for 8vance that can be deployed in organisations to find the best suited startups within their current strategic course. The model consists of a set of features that are divided by classes. Every class has its own function and every feature has its own underlying sub features. Within the class of matching features, both the set of features and the sub features need to be weighted in order to match startups with a selected strategy. Then, the weighted features will get a hierarchical order using the AHP algorithm. Consequently, this will result in a vector which will be multiplied by the weighting of the sub features. The next step is to scout startups. This is done by providing them with an objective quantification based on their current marketing characteristics. This quantification has to be done for the available sub features. Then, the final matching result can be calculated by matching the weighting results with the objective description of every startup.

For this model, 8vance already has an interested party. TIC is willing to act as first launching costumer in order to be the first organisation using this matching technology. It was striking to see that, during the modification phase, we learned that initially, no additional features are needed to make the generic model more specific for TIC. However, modification can be important to ensure that the model fits a corporation's chosen strategies. During the Validation phase, we learned that different strategies need different relative weightings and will result in different rankings of the selected startups. The validation phase showed promising results concerning the quality of the Strategy Matching approach developed during this study. However, the approach needs to be validated more comprehensive by subjecting it to additional strategies. This can be done during future studies. Lastly, a theory has been proposed which has elements to define similarity, distance, and has a method to determine the distance from similarity.

7.2 Further research

In this paragraph, we will discuss different subjects. This concerns subjects that need to be looked at during follow up studies.

Data gathering – interpretation and automation

During this study, we noticed that the most time consuming part was data gathering. Different features need different data which needs to be obtained from various resources. Bear in mind that this data needs to be obtained for every startup. Therefore, this process needs to be automated. Automating the data gathering process does not only mean extracting data. The interpretation of this data also needs to be included since not all the data that is being extracted, is numerical data.

Fine tuning the relative weighting

When using this model for Strategy Matching, the relative weighting process is one of the most important processes. This process needs to be run through iteratively as shown in the 'Validation' chapter in order to fine tune the relative weighting so it corresponds accurately with the selected strategy. During this fine tuning phase, we also recommend to pursue an as high as possible transitivity since this also will contribute in a higher quality of the relative weighting. A higher transitivity can be achieved by reducing the consistency ratio[6].

---

[6] A consistency ratio of 0% means a 100% transitivity.

As described in chapter 4.6, using AHP, the λ_{max} calculation is necessary if you want to derive the CI and CR. During this calculation procedure, a hint can be obtained indicating the need to improve the comparison matrix.

Lastly, we recommend to extend the validation phase with at least two additional TIC strategies. For this extension, we already have asked the TIC stakeholder to provide us with two strategies including the concerning intuitive ranking. The selected strategies, Digital Services and Networked Cameras, are listed in tables 7.1 and 7.2 respectively. They are provided with a brief description. Note that when using these strategies for future research, the TIC stakeholder has to be asked to provide these strategies with transitive feature weightings (comparison matrix) together with the relative weightings of the sub features.

Digital Services strategy

To provide insight into this strategy, we have decided to give a general description of Digital Service.

"A Digital Service is a service that has been entirely automated and which is controlled by the Customer of the Service, for example, as an "app" on a mobile phone or tablet PC. Furthermore, a Digital Service is usually an online service or it contains a significant online component. For example, a Digital Service may use information from a separate computer system or another Digital Service accessed in real time through the internet or an alternative network. Two or more Digital Services can be combined to produce a single, more powerful Digital Service for a Customer"[7].

Startups	Ranking	Matching %
EyeEm	1	80,00%
Orbeus	2	75,00%
Withings	2	75,00%
Narrative	4	70,00%
Canary	5	65,00%
Frontback	5	65,00%
Meerkat	5	65,00%
Storehouse	5	65,00%
Evercam	9	60,00%
Triggertrap	9	60,00%
Podo	11	50,00%

Table 7.1: Intuitive ranking - Digital Services

Networked Cameras

Originally, network cameras were used for surveillance purposes in businesses and public organizations. Nowadays, the target group of these cameras has widened. Consumers also are showing an increasing interest in such technologies. This seems to have emerged a shift in the purpose for which such cameras were used originally. After all, this new target has different requirements than the original target group, and this means that a different approach is necessary.

[7] http://esmarchitecture.com/key-concepts/business-it-digital-services.html

For example, besides functionality, also looks/design will play an important role in the assessment of these cameras.

TIC also wants to address this target group and is looking for startups that have the potential to be successful within this market.

Startups	Ranking	Matching %
Evercam	1	75,00%
Withings	1	75,00%
Canary	3	70,00%
Narrative	4	65,00%
Orbeus	5	64,00%
Meerkat	6	60,00%
Triggertrap	6	60,00%
Podo	8	55,00%
EyeEm	9	45,00%
Frontback	10	40,00%
Storehouse	11	35,00%

Table 7.2: Intuitive ranking - Networked Cameras

Additional features

During the relative weighting phase of this study, the TIC stakeholder came to the conclusion that there are a number of additional features that are considered interesting. Features that could mainly be classified as matching features together with one filtering feature. The concerning features are listed in tables 7.3 to 7.6. To determine whether these features are of any adding value, additional research is necessary.

Classification	Feature	Sub features		
Matching	Application	Advertising	Medical	Mapping
		Agriculture	Navigation	Robot Vision
		Architecture	Photography	Smart Cities
		Art Conservation	Projection	Storytelling
		Astrology	Quality inspection	Surveillance & Monitoring
		Entertainment	Remote Sensing	Publishing
		Biology	Retail	
		Engineering	Manufacturing	

Table 7.3: Application feature - Matching feature

Classification	Feature	Sub features		
Matching	Imaging value	Image analytics	Visual communication	Image generation
		Image security	Mapping	Image broadcast
		Image management	Visual literacy	2D
		Image sharing	Video Surveillance	3D
		Visual advertising	Image compression	
		Image processing	Augmentation	
		Data visualization	Image reconstruction	
		Visual commerce	Image rendering	

Table 7.4: Imaging value feature- Matching feature

Classification	Feature	Sub features
Matching	Imaging type	Processing
		Capture
		Sharing
		Editing
		Storage
		Reconstruction
		Generation
		Display

Table 7.5: Imaging type feature - Matching feature

Classification	Feature	Sub features
Filtering	Imaging value chain	Service Providers
		System Integrators & Consultants
		Distributors
		End users
		Manufacturers

Table 7.6: Imaging value chain – Filtering feature

Supply as a market in strategy market

In this research we focused on the demanding party, the corporation interested in startup adoption. In this scenario, no supplying parties are offering themselves for adoption, simply because no research has been done to determine whether they are interested in this matching process. Therefore, it can be interesting to examine the extent to which startups are interested in takeover. This way, 8vance can find out whether this matching process is also interesting for the startup market and possibly can provide for a service.

The missing data scenario

When providing startups with their sub feature values in the startup scouting process, various data sources are being addressed. This dependency increases the probability of scenarios where startups cannot be provided with all the necessary data. This raises the following question, "To what extent does the completeness of the matching feature values play a role in the quality of the matching result?" Therefore, it may be interesting to study the possibilities to provide every matching result with a matching accuracy. This matching accuracy indicates the level of precision of the matching result based on the completeness of the gathered matching data.

7.3 Final recap

The contributions of this thesis can be summarized as follows:

1. An object model has been developed that allows for the description of real life objects by attribute-value pairs. A special feature of this model is the possibility of decomposition. This makes it possible to introduce levels of abstraction in the description. Attributes may have assigned a weight. We also proposed a recursive similarity function to evaluate the similarity of such objects.

2. The AHP method was proposed to assign weights to attributes. This method provides a simple intuitive method based on comparing attributes pairwise. Then AHP checks consistency. The attribute weights then are computed as the eigenvector of the constructed attribute comparison matrix.

3. The clustering method used by 8vance requires the objects to be described by vectors and a distance measure rather than a similarity measure. We discussed a general strategy to convert similarity into distance.

4. The object model was applied in a concrete problem situation. The application showed that the introduction of abstraction levels made it more easy to define and understand the model. The resulting model was applied in the 8vance context, to generate and the results were promising.

Finally, in the further research section, we have added various recommendations to help further research with a thrifty start.

BIBLIOGRAPHY

[1] S. L. a. Y. K. Ronald R. Coifman, “Data Fusion and Multicue Data Matching by Diffusion Maps,” IEEE Computer Society, 2006.

[2] p. d. i. T. v. d. Weide, *Lecture notes,* Nijmegen: Radboud universiteit, 2015.

[3] E. F. Krause, Taxicab Geometry, Dover: Dover Science Books, 1987.

[4] K. P. Thomas L. Saaty, Group Decision Making: Drawing Out and Reconciling Differences, Pittsburgh: RWS Publications, 2007.

[5] NASA, “Technology Readiness Level,” 12 10 2012. [Online]. Available: https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html. [Accessed 03 2015].

[6] Gartner, “Hype Cycles 2015 Research Report | Gartner Inc.,” 2015. [Online]. Available: http://www.gartner.com/technology/research/hype-cycles/. [Accessed 04 2015].

[7] P. Konstantinos, “AHP_Lesson_1,” [Online]. Available: http://rad.ihu.edu.gr/fileadmin/labsfiles/decision_support_systems/lessons/ahp/AHP_Lesson_1.pdf. [Accessed 15 10 2015].

[8] T. L. Saaty, The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation, United States: Mcgraw-Hill, 1980.

[9] J. Polpinij, “Ontology-based knowledge discovery from unstructured and semi-structured text,” School of Computer Science and Software Engineering - University of Wollongong, Wollongong, Australia, 2014.

[10] N. Z. a. Z. Z. Christoph F. Eick, “Supervised Clustering – Algorithms and Benefits,” in *Tools with Artificial Intelligence, 2004. ICTAI*, Houston, Texas, 2004.

[11] J. A. Bullinaria, *Self Organizing Maps: Fundamentals,* Birmingham: University of Birmingham, 2004.

[12] M. S. V. K. Pang-Ning Tan, “Introduction to Data Mining,” Pearson Education, US, 2005.

[13] T. T. Asfaw, Service discovery in eHealth, Nijmegen: SIKS dissertation series ; 2014-33 (http://repository.ubn.ru.nl/bitstream/handle/2066/131341/131341.pdf?sequence=1), 2014.

APPENDIX A

Appendix A is the taxonomy of the complete set of features. Here, the three classifications are indicated clearly. Moreover, the additional features provided by TIC to make the generic model more specific for their strategies, are also included.

Taxonomy

The complete taxonomy is divided into multiple fragments to prevent the overview from being lost. Every fragment is provided with a brief description. Lastly, a legend is included to explain the frame types used in this taxonomy.

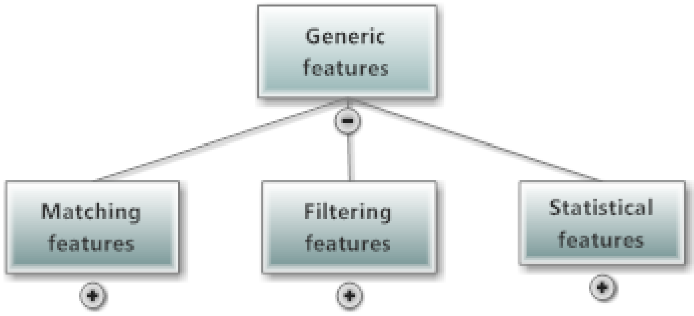

Figure B.1: Feature classes overview.

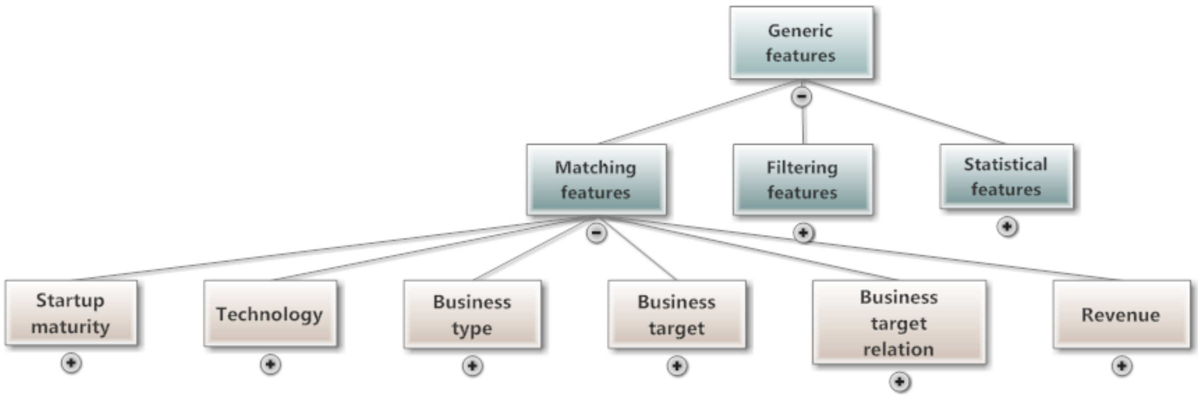

Figure B.2: Matching class, features overview.

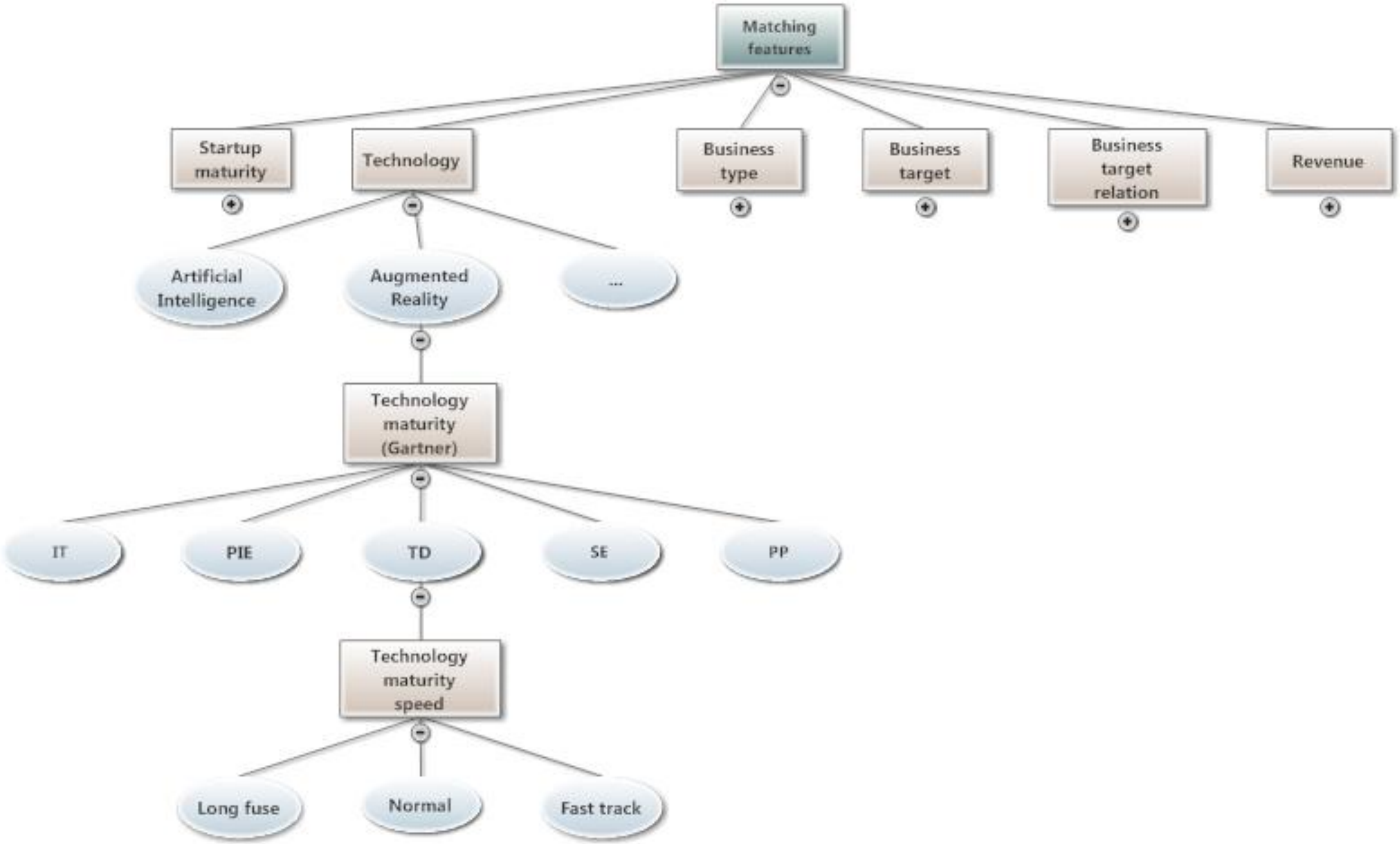

Figure B.3: Technology feature zoom-in.

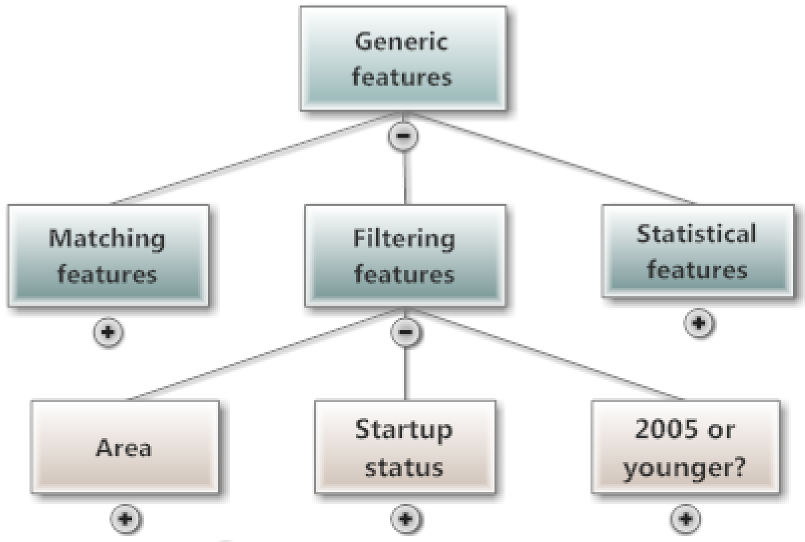

Figure B.4: Filtering class, features overview.

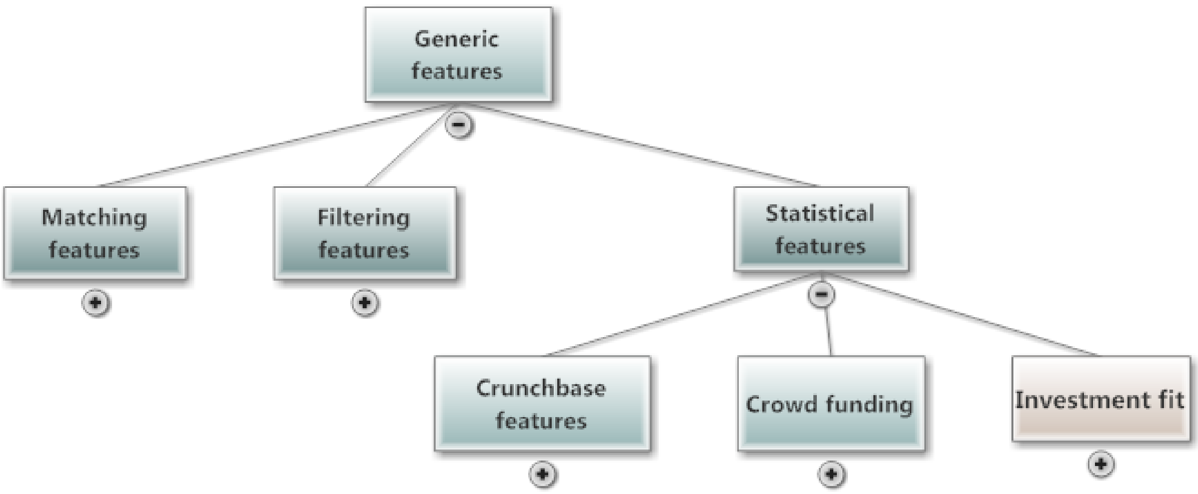

Figure B.5: Statistical class, features overview.

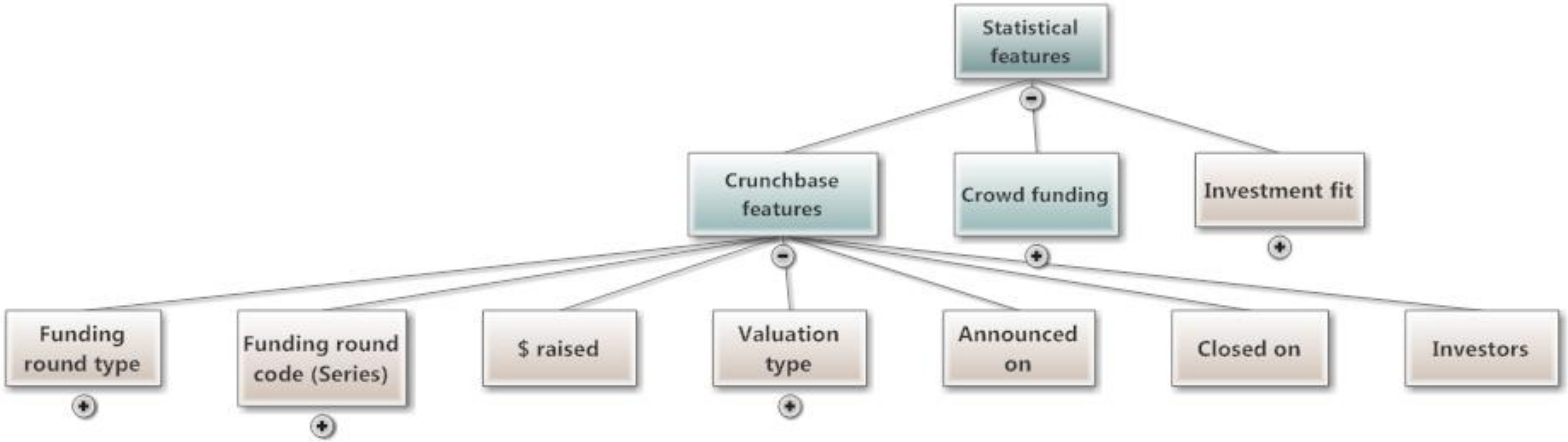

Figure B.6: Crunchbase features overview.

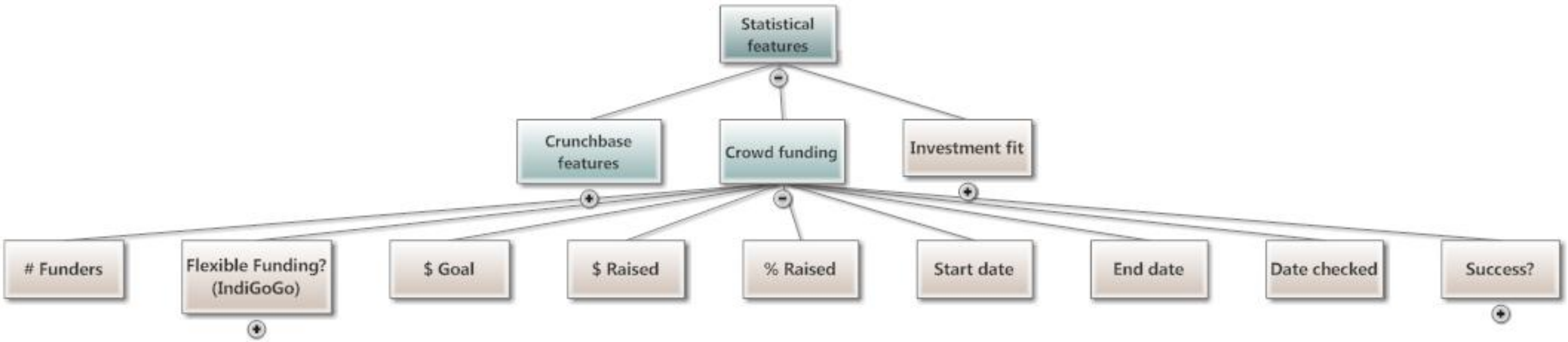

Figure B.7: Crowd funding features overview.

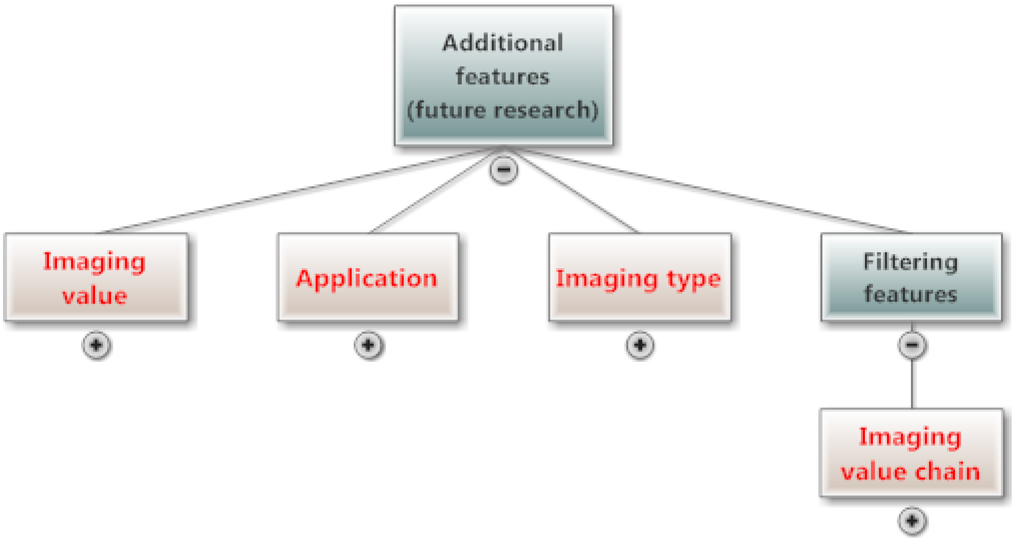

Figure B.8: TIC specific features, three matching features and one filtering feature.

Frame type	Description
	Main frame, used to designate the feature classes or to group a set of features.
	Feature frame. Note that when the text is coloured red, it concerns features that need to be examined in a follow-up study.
	Sub feature frame.

Table B.1: Taxonomy legend