Finding implicitly related items based on semantic similarities and metadata in a non-hierarchical network of documents

Authors

R. van Ginkel | J. Peters | L. Weerts

Project commissioned by *Perceptum B.V.*

Supervisors

Academic Supervisor	Raquel Fernandez
Company Supervisor	Robrecht Jurriaans
	Sander Latour
	Wijnand Baretta

June 24, 2014
Universiteit van Amsterdam

Abstract

Contents

1 Introduction

This report describes the results of the Second Year's project for the Perceptum team. The project focused on creating a *document link recommender system* to the StarFish website.

StarFish, one of the products of Perceptum, is a website that aims to share knowledge about the education domain by means of a connected graph. People from all around the world should get access to this knowledge graph in a simple, personalized manner. The nodes in this graph are documents and they are connected with links. These documents can be of all sorts of types - e.g. a good practice, information, a question. Each document has a set of tags associated with it, which describe the different aspects of educational innovation. StarFish is community-driven: both the content of the documents as the links between documents are determined by the users of StarFish. The drawback of a community-driven knowledge graph is that not all the users have complete knowledge the entire document base. Therefore, many links will not be made because the users are unaware of the documents that they could link to. A possible solution could be to make use of administrators, which can devote more time in getting to know all the documents. However, that approach has two main drawbacks. First of all, this would mean that some central authority determines whether or not two documents should be linked. This is not in line with the idea of a community-driven knowledge base. Secondly, if the knowledge base grows even further, it becomes impossible for an administrator to keep track of all documents. Imagine one person having to link all pages on Wikipedia - an impossible job.

In order to overcome the problem of linking documents in a large knowledge base, this process should be automated. This project focuses on the automatization of connecting documents within Starfish. Though ideally these connections should be made completely automatic, a first step would be to create a recommendation system. When a user adds a new document, he or she can choose from a list of proposed documents the documents he or she deems relevant. This means that the recommender system does not have to work perfect, but should work reasonably well enough. Defining 'well enough', however, is also a part of this project. Thus, the product vision of the system can be described by the following:

Product vision:

For	StarFish users
who	search for and edit knowledge in starfish
the	starfish document linker
is	a starfish core system addition
that	finds related documents
unlike	moderated or individual linking
our	product uses algorithms and data to suggest document links

Within the time span of this project multiple ways of recommending links between documents have been explored. The results of these explorations will be discussed in this report.

2 Domain

The starfish current Starfish document graph contains 240 documents which are directionally linked. Each document in this graph is of one of the following types: Information, Glossary, Question, Good Practice, Project, Person or Event. Documents have an author, title, text, tags

and links. Some document types have different optional fields like 'name' for person and 'headline' for good practices and projects.

Glossaries are special types of documents, as they are description for tags. Because different groups use alternative names to describe concepts, tags may be aliases. Of the 210 tags the current system contains, x unique tag concepts can be distinguished of which x% has a glossary. On average a document has x tags, x outgoing links, and x incoming links.

These numbers and properties of the system give some insight into the current state of the dataset and possible solutions. Due to the nature of the directional links, semantical document analysis will probably not be enough to correctly propose links. A novel Starfish specific method could be combined with this well known symmetrical measure for optimal results. Starfish mostly contains documents in English, but a small part of the data is in Dutch. This language diversity and the fact that documents can also be non-textual like images or videos indicate a system that is not purely based on text will perform better.

3 Product overview

The product created in this project is a python program takes a set of documents and a new document and returns the subset of documents that should be linked with the new document. For this, a descriptor-based approach was used, which consists of three steps. First, each of the documents is transformed into a descriptor: a vector containing numerical values that in some way describes the document (hence the term 'descriptor'). Creating these descriptors is not trivial and during the project several techniques have been explored. Secondly, a ranking is made of all documents based on the similarity of the document descriptors and the descriptor of the new added document. To compare the descriptors the Nearest Neighbour algorithm was implemented, including five different distance metrics that determine how near two vectors are. Thirdly, an algorithm chooses the proper amount of proposed links that must be returned.

```
python documentlinker.py -vectorizer <vectorizername>
-distance <distance metric>
-bayes <true/false>
-threshold <0..1>
```

We will now discuss each of these parameters, since these will give more insight into the approach that was chosen to solve the problem. For the performance of the different parameters we refer to the evaluation section.

3.1 Vectorizer

The first step is to create document descriptors, which is done by algorithms that we call *vectorizers*. Two main paths have been explored: transformation based on text and transformation based on tags.

3.1.1 Text-based transformation

Textvectorizer The text-based vectorizers use the textual content of the documents and are therefore generally applicable to knowledge bases that contain text-based documents. The textual content is first transformed into a *bag of words*. Then, based on all the documents in the knowledge base, the *TF-IDF* value is calculated for each of the words in the bag of words. *TF-IDF* stands for Term Frequency-Inverse Document Frequency and represents the importance of a word to a document in a bigger set of documents. Thus, the document

descriptor consists of a vector with all TF-IDF values for that document of all words in the corpus.

Weighted_textvectorizer The weighted textvectorizer is implemented as an extension of the textvectorizer. First, all descriptors of the documents are calculated similarly as in the textvectorizer. Then each document descriptor is increased with the sum of the descriptors of the document it is linked to itself, decreased by some weight. This captures the idea that if a new document resembles some of the documents that are linked to one particular document, it is more likely to be linked to this particular document.

3.1.2 Tag-based transformations

Simple_tag_similarity The tag-based transformations are more StarFish specific, since they make use of the tags that are assigned to the documents. A tag is a keyword that describes a topic/term that is important for that document. For example, 'Online Support and Online Assessment for Teaching and Learning Chemistry' is tagged with 'chemistry', 'e-learning' and 'assessment'. The simple tag similarity vectorizer creates a vector where each value indicates whether or not one particular tag is assigned to the document.

Tag_smoothing The tag smoothing vectorizer uses the co-occurence of tags in estimating document similarity. Even though tags might not co-occur on any document in the data set, they can still provide information about each other. For example, the dataset exists of documents with associated tags like $\{\{t_1, t_2\}, \{t_1, t_3\}\}$. From the co-occurence it does not follow that t_2 and t_3 are related, however by transitivity with t_1 we want to create a small implicit link between t_2 and t_3. The tag smoothing method does this based on work from Zhou et al. (2011).

Glossaries_of_tags Another way of capturing tag similarity is by using tag Glossaries. A Glossary is a type of document that holds a textual description of a tag. Though Glossaries are documents, they have the special role of describing tags. This also means that a document should not be linked with a Glossary, since it would be better to simply tag the document with the Glossary that describes it. The glossaries can still be used by applying a text-based transformation on the glossaries to indicate the similarity between tags. Thus, glossaries_of_tags can be seen as a hybrid form of the tag and text-based approaches, where the glossary of a tag is turned into a TF-IDF bag of words. The document descriptor consists of the sum of vectors of each of its tags.

Weighted_tag_vectorizer This is an extension of glossaries of tags. In the original glossaries of tags, it is assumed all tags contribute the same amount of information to a document's links. In practice some tags provide more information than others. If a certain tag is on nearly all documents in the dataset, it does not provide a lot of insight into linking new documents. In contrast a tag which is only attached to a small subset of documents is much more informative. The weighted tag vectorizer creates descriptors by summing the tag vectors with a weight based on the frequency of that tag in the dataset.

3.2 Distance

The Nearest Neighbour algorithm loops through all available document descriptors and compares these with the descriptor of the new document. The closer a descriptor is to the descriptor of the newly added document, the higher its ranking will be. The distance metrics define the closeness of the descriptors - a lower distance means a closer relation. The following were implemented:

Eucledian

Cosine

Bhattacharyya

Correlation

Intersection

3.3 StarFish specific adaptations: Bayesian weighting

Both the tag-based and text-based approaches uses some kind of 'semantic similarity' - the similarity of tags or text. However, except for the weighted text vectorizers, no information about possible links is used. For example, the text on a person's profile might be similar to other persons, but within Starfish a person is almost never linked to another person. In the Bayesian weighted vectorizer this is captured by weighting the vectors with the probability that two documents are linked together given their tags:

$$P(D_a \rightarrow D_b|t)$$

Thus, the weight of a tag within a vector is equal to the chance that given this particular vector, a document of type a (the type of the newly added document) and a document of type b (equal to the type of proposed link) are linked together.

3.4 Threshold value

The next step in the pipeline is to determine how many of the nearest neighbours should be returned. Depending on the application of the starfish document linker, the desired number might vary. If one wants to immediately link the results, the certainty for relatedness should be high. If the links are presented to a user which can approve or reject them, the relatedness may be lower. Currently, this is configurable by setting the threshold parameter between 0 and 1. Zero will only return the closest document, 1 will return almost all. After exploration of the dataset the default value is 0.3, which roughly returns the same amount of links which is currently average for Starfish.

4 Method

4.1 Text vectorization

4.1.1 Textvectorizer

The first set of vectorizers focuses on the texts of the documents. The *textvectorizer* is a very generic approach that can be used on any corpus of textual documents. In the StarFish context, we define 'content' as the title and text-fields of a document. The only exception on this are Persons, of which we wil use the xx and xx.

The textvectorizer makes use of a bag of words representation. If two documents cover the same subject(s), they are likely to contain similar keywords. To capture this similarity, the documents can be transformed into a list of all words that are present within that text. Instead of counting the frequency of each word within a document, the more sophisticated Term Frequency-Inversed Document Frequency value was used. TF-IDF is a statistic that reflects the importance of a word

in a document within a corpus by inducing a trade off between the term frequency, the number of times a word appears in a document, and the inverse document frequency, the inverse of how often a word is used in the entire corpus, see formula xx.

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d) : w \in d\}}$$
$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$
$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

Thus, words that are generally common, get a lower value. The TF-IDF was calculated using scikit-learn (Pedregosa et al., 2011). Though the TF-IDF values of words that are used very often should be low, common words such as 'and', 'or' and 'of' are still present in the vectors. This could be caused by the different types of documents. For example, a Question often structured in a less complex way than a Project description. Adding a standard English stopword list to the vectorizer improved it's performance

Table x shows the performance of the textvectorizer on the data set. It shows that of all (valid) documents, about 20% of the links that were returned were correct. If the results are split into types of documents, it can be seen that the textvectorizer performs best in questions. This can be explained by the nature of questions: they often contain important keywords that indicate the subjects the question is about. If a document is a Person, on the other hand, only xx percent of the proposed links is correct. Table x shows a possible explanation of this phenomenon. The textvectorizer often proposes other Persons if a new document is a Person. However, within StarFish a Person is rarely linked to another Person. This implies that a more StarFish specific approach that does make use of links is more reasonable.

4.1.2 Weighted textvectorizer

The weighted text vectorizer is an extension of the textvectorizer that takes into account the links of the proposed documents. The vectors of the links of a document are added with some weight to the vectors of the documents themselves. Intuitively, this would add semantic information about a document based on it's links. For example, a Person is likely to write other documents about his or her subjects of expertise. Knowing not only the biography of a Person, but also the content he or she has added to StarFish, gives a more complete image of what documents could be related to that Person.

The vectors of links of a document are added in a recursive way, where documents that are linked directly have a higher weight than documents that are linked transitively. The algorithm is displayed in figure xx.

The weighted text vectorizer performs better than the normal text vectorizer.

4.1.3 Text-based approach limitations

Overall, both textvectorizers are slow in performance even though the corpus is small. Additionally, the the bag-of-words approach imposes a few limitations on the document linker. Firstly, it performs bad when different languages are used. Figure x shows the vectors of three texts when an English text is used, combined with an English proposed document and a Dutch proposed document. The English and Dutch vectors have only little words in common - luckily the keywords are in this case English, but if they are not this is a problem that cannot be overcome by simply looking at texts. Secondly, the current StarFish network consists of mainly textual content.

However, in the future this is likely to be extended with images, videos and other non-textual content. These sources should then somehow be converted to text.

4.2 Tag vectorization

4.2.1 Simple tag vectorizer

The tag-based approach is more StarFish specific than the text-based approach, since it depends on the tags that are available in StarFish. The tags on StarFish are added by the users themselves, so offer a human-based vision on what a document is really about.

The simple tag vectorizer is a very straight forward implementation of the idea of using tags. The vectors of this transformation consist of a binary list that tells whether or not a tag is attached to the document.

Due to it's simplicity, the simple tag vectorizer is very fast. It's performance, as shown in table xx, is about 24% precision. Both Question documents and Person documents perform quite bad. This can be explained by the fact that half of both Questions and Persons have zero tags. Obviously, the simple tag vectorizer cannot deal with such documents. In fact, almost all other Questions have only one tag. Since the simple tag vectorizer compares vectors, it wil prefer documents that also have only that particular tag, which makes it sensitive to attaching Questions to Questions. Something similar seems to happen with Persons, of which 45% of the connections are with other Persons. Apparently, persons with similar expertices are tagged similarly. However, as mentioned with the tag vectorizer, in StarFish persons almost never refer to other persons. Moreover, if a document is badly labeled this can also induce problems. For example, take the question 'Is there an English version in Tentamenlade', tagged with 'ToetsenEnToetsgestuurdleren'. The proposed links are visible in table xx, which shows that the three proposed questions all have the tag 'ToetsEnToetsGestuudLeren'. However, if the question was tagged with the tag 'Tentamenlade', which seems very reasonable given the proposed question, the false negatives would likely be returned correctly by the system. Good practices, events and projects perform significantly better with 75,6%, 73,0% and 35,8%. These are often thoroughly tagged, as shown by document xx, which has a rich set of tags. However, these document types only entail 3.2%, 2.7% and 5.4% of the total amount of documents respectively, which explains why the total performance is still stuck at 24.8%.

```
False Positives:
- Wat is het verschil tussen Learning Analytics en TTL (Question)
- Formatieve meerkeuze toetsen om begin kennisniveau te toetsen (Good Practice)
- De toetscyclus (Information)

True Positives:
- Tentamenlade2.5 (Project)

False Negatives:
- Tentamenlade Natuurkunde (Information)
- Hoe kan ik inloggen in Tentamenlade? (Question)
- Hoe kan ik inloggen in Tentamenlade? (Question)
```

4.2.2 Tag smoothing

The tag smoothing vectorizers creates descriptors based on the tag set of a document. A tag co-occurs with other tags in a document, we assume documents with similar tags should be linked

in Starfish. Let the frequency of occurrence with other tags across the dataset will form a vector for each tag. The descriptor for a document is then created by combining the occurrence vectors for all the document's tags. Now documents with tags that occur together will be seen as similar.

There are two reasons why one would like to smooth the tag co-occurences. Firstly, a problem for this is that tags must occur together before the algorithms works properly. The starfish dataset contains a lot of tags that only occur with a small frequency, which means the tag occurrence vector will contain many zeros. This makes the algorithm perform bad with little data. Secondly, two tags can describe the same concept and be connected to that concept through a common co-occurence with another tag. Whilst they describe the same concept and are connected to that, they are not directly linked together. Therefore it seems feasible to perform some sort of smoothing on the co-occurences of tags.

Zhou et al. (2011) proposed a method to cluster web documents based on tag set similarity. This is based on a similarity between two tags as a relation between the frequency these tags occur separate and together, as described in equation 1. To smooth these similarities between tags, a tag similarity matrix $\mathcal{C}$ is constructed. Each entry $c_{i,j}$ in this matrix can be viewed as the angle $\theta_{i,j}$ between two unknown vectors v_i and v_j. These vectors cover both explicit similarity and implicit similarity (Park et al., 2010). This transfers the problem to find a set of linearly independent vectors $\{v_1, v_2, \ldots, v_n\}$ for which for all $v_i \cdot v_j = \cos\theta_{i,j}$. One must find a matrix $\mathcal{V}$ for which $V^T V = C$. This can be done by orthogonal triangularization on $\mathcal{C}$ for which Zhou et al. introduces a modified Cholesky transform.

$$s_{i,j} = \frac{f_{i,j}}{f_i + f_j - f_{i,j}} \tag{1}$$

Evaluation in cijfers hier

In the current implementation this vectorizer is relatively slow. In practice the similarity matrix can be pre calculated and updated in batches. Due to the transform on the tag similartity matrix, it is very hard to determine which tag occurrences contributed to the document similarity and why some recommendations are made. It does not seem to perform much better than the regular bag of words tag descriptor, in Zhou et al. the algorithm only starts performing significantly better when it is presented with more tags.

4.2.3 Glossaries of tags

- Motivation: find underlying network between tags by using their glossaries
- Implementation: hybrid form of text and tag vectorizer
- Evaluation: does not work well (find out why!)

4.3 Distance metrics

- Eucledian
- Cosine
- Bhattacharyya
- Correlation
- Intersection

4.4 Bayesian weighting

Explain motivation, implementation and short results

4.5 Thresholds

Explan motivation, implementation and short results

5 Experiments

Overview of performance over the entire pipeline

6 Conclusion

- Which vectorizers do not work
- Which vectorizers do work
- How well does the bayesian layer perform
- How well does the thresholds perform

7 Future Work & Recommendations

Latent dirichlet allocation because tags in dataset are not so good.
Also create incoming links

References

Park, J., Choi, B.-C., and Kim, K. (2010). A vector space approach to tag cloud similarity ranking. *Information Processing Letters*, 110(12):489–496.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Zhou, J., Nie, X., Qin, L., and Zhu, J. (2011). Web clustering based on tag set similarity. *Journal of computers*, 6(1):59–66.