THE WORLD

ACCORDING

TO THE

WE

B

DOUWE M. OSINGA

ERNST C. WIT

囗

1. Introduction

Much has been said about the World Wide Web.[1] It was developed at CERN for high-particle physics to share information among researchers and it is currently seen as an information resource on a larger scale. Google and other search engines in this "big telephone directory"-paradigm are simply more or less efficient tools to find pieces of specific information among many pieces of unspecific information.

Whereas in the early days of the World Wide Web, this paradigm might have been an appropriate metaphor, this is no longer the case. To view the World Wide Web as an archive in which some independent pieces of information live is to think of a medieval concept of substance, namely as the unspecific "carrier" of secondary qualities such as colour and temperature.

Just as this paradigm collapsed in favour of a more holistic approach to matter as molecular that essentially defines its substance and qualities simultaneously, so it is more appropriate to think of the World Wide Web as ecosystems of information and Google, Yahoo! and other search engines as alternative interpretations thereof. They do not just disclose information, but structure it. As the size of the World Wide Web has ballooned, in almost all cases there is no independent substratum of the information on the Web.

Although this fact can lead to (less) amusing details such as invented personalities and unsubstantiated myths, more importantly it suggest that the way the Web is structuring information is worth

studying. The aim of this paper is to show a particular implementation of a mapping algorithm of symmetric relationships between objects within the Web based on AltaVista's ordering thereof. In particular, we shall use the example of mapping the countries of the world based on an information-metric. Certain symmetric relationships, such as "war" might alter this metric and lead to new maps. Multi-dimensional scaling techniques such as Sammon Mapping are particularly useful in projecting the objects from a high-dimensional information-space to a two-dimensional map-space.

2. A measure of Web-dissimilarity

The world is a 3-dimensional object, in which countries, or at least their gravitational centres, can be represented via a 3-dimensional vector. A 2-dimensional map of the world is an attempt to display these countries in such a way with respect to each other in order to preserve a particular property of the underlying 3-dimensional reality. The Mercator projection dates to 1569 and was used for navigation because straight lines on this projection represent lines of constant bearing. Albers Equal-Area Conic Projection projection, as its name implies, maintains equality of area over the map.

For mapping the information world, the situation is both more and less complicated. Rather than objects in a 3-dimensional space, countries in the information space are objects in an extremely high-dimensional space. As every web-page is a dimension, the information universe is, currently, approximately 3 billion dimensional and growing every day. On the other hand, a country is merely a point in this high-dimensional space, which makes the issue of invariance much simpler. Rather than preserving area or straight lines, the most interesting form of invariance is that of *distance* or *dissimilarity* between the countries.

It is therefore important to define a dissimilarity measure in the 3-billion dimensional information space. We start by proposing the following similarity measure *s* for the similarity between countries *a* and *b*:

$$s(a, b) = \frac{\text{Number of pages with both } a \text{ and } b \text{ on it}}{\sqrt{} \text{ (Number of pages with } a \text{ on it} \times \text{Number of pages with } b \text{ on it)}}$$

This is a form of correlation similarity:

· The strongest similarity is equal to 1, which means that on every page on which *a* is present, also *b* is present and vice versa.

· The measure is scale invariant: if the internet would double in size by simply duplicating all existing pages, the similarity *s(a, b)* would remain the same.

There are, however, a few distinct differences between correlation and this information similarity:

· Correlation ranges from −1 to +1, whereas the measure *s(a, b)* ranges between 0 and +1.

· If no *a* is not present of any of the pages of *b*, then the measure *s(a, b)* equals 0. This means that *a* is really unrelated to *b*.

The measure s(a, b) is easily transformed into a dissimilarity measure d:

$$d(a, b) = 1 - s(a, b)$$

A dissimilarity of 0 corresponds with the fact that *a* always appears with *b* on the same page and vice versa, whereas a dissimilarity of 1 corresponds to the case in which *a* never appears on the same page as *b*. Despite its intuitive appearance, *d* is not a metric as the triangular inequality does not apply.

The advantage of *s(.,.)* and *d(.,.)* is that they can easily be extended to measure the similarity of two countries, or objects in general, in the presence of a symmetric relationship *R*. For example, if *R* is the term "war", then it is sensible to define the symmetric relationship:

$$a \, R \, b,$$

which means that country *a* is at war with country *b* and vice versa. Putting a similarity measure on this relationship means assessing the *relevance* of this relationship. We extend the similarity measure for *a* and *b* in relationship with *R* as follows:

$$s(a, b \mid R) = \frac{\text{Number of pages with } a, b \text{ and } R \text{ on it}}{\sqrt{(\text{Number of pages with } a \text{ and } R \times \text{Number of pages with } b \text{ and R})}}$$

Depending on the relational term R, the similarity measure $s(., . \mid R)$ can be interpreted as a *probability*, a *weight* or as a *prevalence*.

3. **Methodology**

The result of applying the dissimilarity measure to a set of n objects or countries is a $n \times n$ dissimilarity matrix X with zeroes on the diagonal. Although these n objects live in a p-dimensional space with p much greater than n ($p \gg n$), it is possible to represent these n objects in a $n-1$ dimensional space without distorting the dissimilarity matrix. However, as n is typically much larger than 2 or 3, this is insufficient for practical visualization purposes.

There exist several dimension reduction techniques, the most famous of which is the Singular Value Decomposition. This method decomposes the distance matrix X into the product of two unitary matrices U and V and one diagonal matrix D,

$$X = U D V^{\mathrm{T}}$$

The diagonal elements of D are positive and can be ranked in order of importance, $d_1 > d_2 > \ldots > d_N$. By considering only the first two columns of U and V, one can "reconstruct" the best possible 2-dimensional reconstruction of X. The first two columns of U represent the coordinates for the 2-dimensional projection of the n countries.

One of the major problems with this approach is that the first and second coordinates are not symmetric: the scaling in each of the coordinates is proportional to the size of the diagonal elements of D. Using the raw coordinates could therefore give a false impression of the distances between objects as the distances in the second coordinate do not have the same interpretation as the distances in the first coordinate. Particular for mapping purposes this is undesirable.

THE WORLD ACCORDING TO SOFTWARE

To overcome this drawback, we propose to use a multiple dimensional scaling technique, known as Sammon mapping[2]. This technique aims to find the best two-dimensional representation that minimizes the "stress" of the distortion from the $n-1$ dimensional distortion-free representation, whereby stress is defined as

$$\text{Stress} = \frac{1}{\sum_{i \neq j} d_{ij}} \sum_{i \neq j} \frac{(d_{ij} - d'_{ij})^2}{d_{ij}},$$

where d_{ij} are the distances in the original high-dimensional space, whereas d'_{ij} are the distances in the reduced, 2-dimensional space.

The original Sammon algorithm treats all nodes as equal, i.e. the stress between any two points in the system has the same impact on the overall stress as any other two points. However, when mapping "war", for example, it might be more important to get the relative distance between Germany and France or between the U.S.A. and the U.K. right, rather than the distance between Costa Rica and Lithuania. Countries have different importance, which is typically defined by their economic and military muscle or their sheer size. When mapping the world of the Web, we need a weight measure of this importance that is defined in terms of the Web itself.

We propose a simple measure of web-importance of any object or country a,

$$w(a) = \sqrt{\text{number of web pages with } a \text{ on it.}}$$

Taking the square root of count data is a natural operation and has many theoretical advantages[3]. It stabilizes the variance, which tends to increase with the number of counts.

By combining the weights of the individual countries we can construct a weight function for all the country pairs, i.e.,

$$w_{ab} = w(a) \times w(b)$$

Given this weight function, we propose a *weighted Sammon algorithm*, which projects the objects in a *k*-dimensional (typically *k=2*) as to minimize the *weighted stress* between the objects, where

$$\text{Weighted Stress} = \frac{1}{\sum_{I \neq J} w^{IJ} \, d^{IJ}} \sum_{I \neq J} w^{IJ} \frac{\left(d^{IJ} - d'^{IJ}\right)^2}{d^{IJ}},$$

where, as before, d^{IJ} and $\ddot{\imath}^{IJ}$ are the distances in the original high-dimensional space and in the reduced, *k*-dimensional space, respectively. It should be noted that adding weights to the optimality criteria is not the same as considering the unnormalized similarities. If the internet were to expand by copying each pages twice, the solution to minimizing the weighted stress would be the same, whereas considering the unnormalized (dis)similarity would typically result in a different solution.

The result of this weighted Sammon algorithm is that in most relationship-maps there is a superstructure between the more important nodes, which are also the centres for local clusters. For example, the fact that there are few pages about Costa Rica and Lithuania together doesn't mean that these countries have to be located far away from each other, just as long as they are well-placed with respect to their respective important neighbours: the US and Russia (most likely).

4. Technical Implementation

The set-up of the mapping programme consists of two parts: (i) there is the harvester program that counts the results for various searches and (ii) there is the stress optimiser that takes the result of the harvester program, calculates the optimal configuration (i.e. the one with the least weighted stress) and that displays this configuration on the screen. Both programs are written in Python, an open source scripting language with a lot of standard support for Internet protocols.

The harvester program sends standard HTTP-requests to a search engine and extracts the number of results from the returned page.

During the development it became clear that Google is not the best search engine for the experiment. Not all co-occurrences of two words on the same page indicate a correlation; some very large pages could contain all kinds of unrelated words, for example in a mailing list overview. Those pages don't usually turn up in search results on Google because of their relative irrelevance, but they do count in the number of hits returned. AltaVista allows users to search for keywords that occur near each other using the keyword NEAR. This construction restricts the results to only the pages where the names of countries occur near to each other and where there is a higher chance of an actual correlation. For this reason, AltaVista was chosen to work with.

5. Findings

The mapping algorithm was applied to the World Wide Web. Figure 1 shows the mapping of the 25 most prevalent countries on the Web according to their web-dissimilarities. The upper part of the map is dominated by the Anglo-Saxon countries; To the right sits Europe, with Germany and France in an ever closer union. The
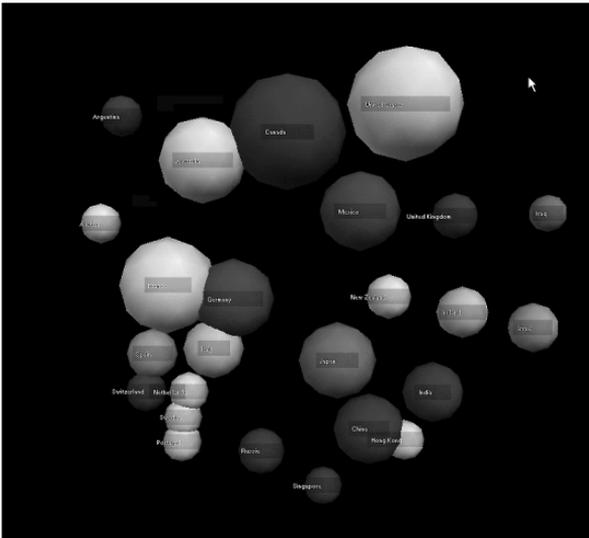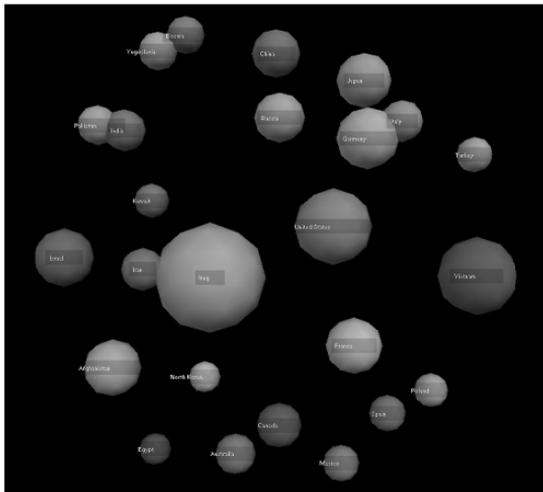
1   Map of the world according to the Web

lower part of the map consists of Asia, with Japan the most central because it is the most westernized nation there. Another interesting feature is the striking absence of most of the African countries. Besides Iraq and Israel, none of them featured in the top 25 countries in the world.

Figure 2, the *world war map*, is quite spread out; so many wars, so little space on this bitmap. The Second World War axis is visible in the right upper corner, as are the "axis of evil" in the lower right corner. Russia, China and Japan form an uneasy triangle of historic conflict. The US is in the middle of every conflict, it seems. Good news for Tony Blair: the United Kingdom is absent from the map. Does Britain not belong to the top 25 aggressive nations in the World? Could this be due to the fact that England, UK and Britain are often considered synonyms and only one term is included in the search?

Figure 3 presents the result of the world map with respect to the term "economy". In the middle reside the powerhouses of the world, the US, Japan and Germany, with China and its two satellites Hong-Kong and Taiwan waiting in the wings. Local economical structures such as Germany-France-Netherlands, Australia-New

2   Map of the world according to the Web with respect to "war"

THE WORLD ACCORDING TO SOFTWARE

Zealand and Mexico-Canada-United States become also clear from this plot. What about the cluster India, Russia and Israel? Maybe their vicinity is explained by the fact they are second world countries that do well on technology.

6. **Extensions**

Mapping is an intuitive mental operation that condenses a lot of information into a practical plot. There are many extensions possible to this general idea. We discuss a few of them here.

6.1. Higher-dimensional mappings and residual stress

The weighted Sammon mapping is easily extended to more than 2 dimensions; in fact, the definition of a Sammon mapping is essentially k-dimensional for any value of k. Although this is not particularly practical for paper representations, standard Python plug-ins make 3D-representations easily accessible to any computer user.

More fundamentally, one can define a measure of the remaining stress in any particular representation of the original $n-1$ dimensional space. The residual stress is defined as
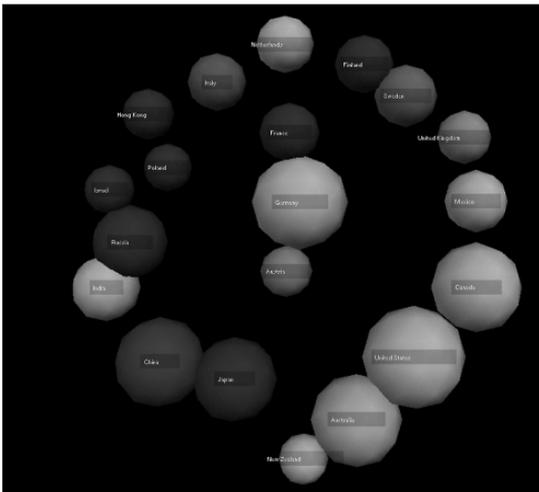
3   Map of the world according to the Web with respect to "economy"

$$\text{Residual Stress} = \frac{1}{\sum_{I \neq J} w^{IJ}} \sum_{I \neq J} w^{IJ} \frac{\left|d^{IJ} - d'^{IJ}\right|}{\max(d^{IJ}, d'^{IJ})} \, ,$$

is a number between 0 and 1, indicating the residual stress in the figure. It is a useful diagnostic for evaluating the success of any particular weighted Sammon mapping.

- The higher the Residual Stress in any mapping, the less suitable the chosen dimensionality is for the specific domain. In other words, by looking at the Residual Stress for a certain domain, we get an indication as to the complexity of the relations within that domain.
- In the three examples above, the Residual Stress is respectively 0.228 for economics, 0.0192 for war and 0.154 for the unqualified map. The fact that the information domain of war is structurally simpler than the other two should surprise no one, for the simple fact that while war is unpleasantly common, the number of countries that have had war is each other is much smaller than the number of countries that trade with each other. Wars typically have a relatively small number of participants.
- That the informational structure of the economy realm is more complex than that of the unqualified map, might seem counter-intuitive, but since the economical relations between countries can hardly be more complex than all relations between countries. But the unqualified map represents the average complexity of relationships between countries, which is dominated by their geographical nearness, which is fundamentally two-dimensional. Countries trade a lot with their neighbours, but are not limited to that.
- If the residual stress is particularly large (> 50%) and if it is impossible to represent a map in a higher dimensional space, it might be advisable to make a Residual weighted Sammon mapping that tries to capture some of the residual stress in the data. The easiest way to achieve this is by rescaling the weights according to

$$w'_{IJ} = w^{IJ} \ \frac{|d^{IJ} - d'_{IJ}|}{\max(d^{IJ}, d'_{IJ})},$$

This means that pairs that were correctly represented in the original weighted Sammon mapping ($|d^{IJ} - i^{IJ}| \approx 0$) get a small weight relative to pairs that were badly represented in that mapping.

6.2. Mapping asymmetric relationships

This mapping tool is particularly suitable for mapping symmetric relationships. The reason lies in the use of spatial vicinity ¾ fundamentally a symmetric relationship ¾ to indicate the strength of a certain relationship. The use of non-symmetric terms is a little less satisfactory. Figure 4, for example, shows the mapping of 25 countries with respect to the keyword "immigration". Although the pic-
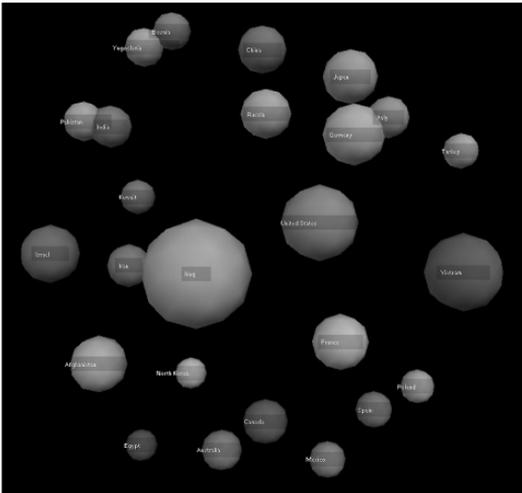

4   Map of the world according to the Web with respect to "immigration". Ireland and Germanys are in the middle as large contributors to American immigration, but nowadays destinations themselves. The big immigration countries keep their distances from the each other and warp the map a bit

ture captures several interesting characteristics about immigration flows in the world, it is essentially unable to describe direction of this immigration.

It would be useful to extend the mapping tool to be able to deal with such asymmetric term. Using arrows of different sizes to indicate an asymmetric distance matrix might be an idea, although spatial vicinity might still be used for a particular function, perhaps to represent the weight matrix, separating the most important countries from each other in order to make the picture as clear as possible.

One way to do this would be by adding extra words to the query, i.e. take as query USA "FROM GERMANY" IMMIGRATION, or combinations with "TO USA" etc. This way the direction of a relation can be determined.

7. Applications

The ideas developed on the billions of pages that make up the Web are a model for reality. For the first time, search engines allow us, using the techniques described above, to mine the collective structures of thought, be it so far only with rather crude

5   Trade map according to the Web

statical tools. As we described in the previous section, many kinds of improvement can be thought of, but the important question is whether the structural analysis of web data yields anything more than pretty pictures.

One of the applications of the practice described here is the direct comparison of the structure of thought and the structure of reality. Take for example trade. Trading relations are relatively easy to measure by looking at the size of trade between two countries relative to the total trade of those countries (or we could relate this to the total trade of either the biggest or the smallest country; note that this asymmetry is very similar to the asymmetry we find when analysing the web). We can then take these trading relations and convert them to a distance matrix and use the distance matrix as input for a Sammon mapping to produce a map of the trading relations in the world. We can compare this map to the map produced by analysing trade as a keyword on Altavista.

At first sight, the maps appear more different than they really are. Remember that for a given data set, a Sammon mapping can produce a multitude of maps that all have the same minimal residual
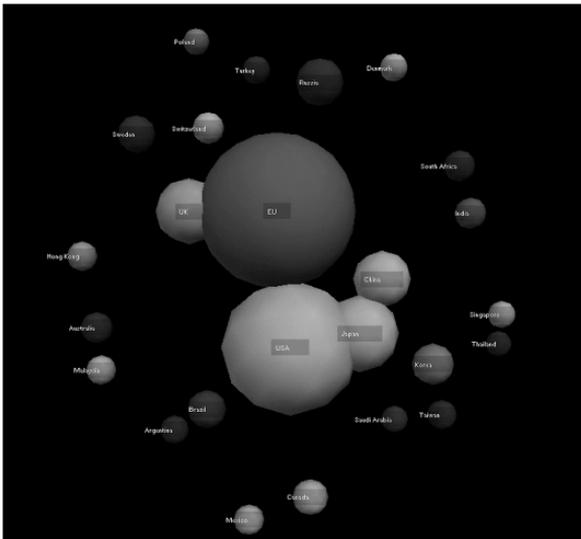
6   Trade Map according to actual trade flows

stress (even if we filter out obvious symmetrical transformations). Structurally, these kind of maps are clusters of related countries that are grouped as well as possible, but between different projections the orientation of these clusters with respect to each other can vary greatly.

Also, since the trade flow data treats the EU as one block, the data seems not very comparable to begin with, but because of the automatic clustering this doesn't make much of a difference — the EU ends up as one block anyway. Moreover, some interesting conclusions can be drawn. What is striking about the actual trade flow picture, is the central axis of the developed trade blocks, Japan, USA, EU and the United Kingdom. The other nations float around this system, with the Asian countries only as a secondary cluster.

In the web picture, weights seem to be more evenly distributed, though this could be partly due to the mapping of trade volumes on number of hits on the web. More interesting is the fact that Israel, Egypt and Iraq appear as major players, while these countries are absent from the actual trade map. At the same time Saudi Arabia and Argentina are missing. The former set of countries seems to be overrepresented in the Internet trade discourse, while the latter is underrepresented.

The second striking thing about the web picture is the fact that the US no longer has a central position, but has been pushed to the side. This can actually mean two things, either the US doesn't have very strong correlations with other blocks in the trade realm as far as the web goes, or the discourse on US and trade is more evenly distributed than others and a lot of talk about trade and the US concerns countries not on the map. The later theory seems more probable. The US has been trying to convince a large number of smaller countries to do a bilateral trading deals, generating a lot of buzz, while the actual trading flows with Japan and the EU are much larger, but also more silent.

THE WORLD ACCORDING TO SOFTWARE

8. Conclusions

We have presented a general method for mapping structures present in the high-dimensional space of the Internet into a visual two- or three-dimensional space. In many ways it shows how the World Wide Web is a reflection of the world as we know it. In other cases, there are subtle differences between the World of the Web and the outside world. As information is easily duplicated and novelty is a considered a valuable attribute of information, we hypothesize that the Web maps tend to reflect people's current preoccupations more than can be accounted for by mere fact. The presence of Iraq in many of the previous maps gives some credence to this hypothesis. One thing that has become clear is that the World Wide Web
is not an amorphous collection of facts, fiction and mere noise.

The Web is guided by its own multifaceted discourse
and guided by the same structures that generate
our world. By analysing the Web, we therefore
get an interesting insight into the
organization of the world
itself
.

1   20 May 2004. http://www.google.com/search?sourceid=
    navclient&ie=UTF-8&oe=UTF-8&q=define%3AWorld+Wide+Web.

2   Sammon, Jr., J.W., *A nonlinear mapping for data structure analysis*,
    IEEE Transactions on Computers, 18:401-409.

3   McCullagh P. and J. A. Nelder, *Generalized Linear Models*, 2nd edition,
    Chapman & Hall, 1989.