Formal Topology and Search Engine

Silvio Valentini
Dipartimento di Matematica Pura ed Applicata
Università di Padova
via G. Belzoni n.7, I–35131 Padova, Italy
silvio@brouwer.math.unipd.it

November 5, 2001

Abstract

Formal topology can be the key tool to create a new kind of search engine for finding information in the web.

1 The need for topology in a search engine

The purpose of a search engine is to find all the documents that can be of interest according to some search criteria (and possibly to order them according to some order relation). There are at least two main approaches in building a search engine: to collect all the documents that satisfy all of the search criteria or/and to exclude all of the documents that do not satisfy such a criteria. In general it is not easy to decide what to exclude and here is where topology can be of some help.

Topology is concerned with a sort of "vague" set theory, that is, a set theory where the elements of a set can be distinguished in a non precise way only by means of their properties. In our case the set we are interested in is the set of all the possible documents (for instance: all the possible books in a library, all the documents in the web ...) and the properties are the particular properties of such a documents (for instance: all the books dealing with calculus, all the text documents containing the word "microsoft" ...).

In order to establish a relation between the documents, that is, the elements/points of the set, and their properties we need a relation linking them: to keep the highest level of generality we can consider the most general kind relation, that is, a binary relation linking any point with all the properties that it satisfies and we can write

$$d \Vdash a$$

to mean that the document d satisfies the property a, where d is an element of the set D of the documents and a is an element of the set P of the properties.

In general no condition is required on the relation $\Vdash$, but sometime some conditions can be of some interest. For instance, one can assume that all the documents posses some property (at least the property of being a document!) which can be formally expressed by saying that for all $d \in D$ there is $a \in P$ such that $d \Vdash a$. Indeed, if we lack of such a condition for some document d we have no search criteria which is going to let us find the document d; so the document d is "not-existing" from our point of view since it is not *visible*.

Other conditions on $\Vdash$ depend on the language that we are allowed to use in expressing the properties. For instance, supposing the set of properties is closed under conjunction, namely, if a and b are properties then also (a and b) is a property, then it is reasonable to assume that if for a document d, both $d \Vdash a$ and $d \Vdash b$ hold, that is, d satisfies both the property a and the property b, then also $d \Vdash a\ \mathsf{and}\ b$ holds.

Other conditions on $\Vdash$ can be proposed as well and they are going to change the search engine that we are going to build, but the two conditions above are interesting for us since they are the only two conditions which are required to speak of topology from a mathematical point of view, that is, to enter in a well-studied mathematical topic (really, the second condition can be weakened a bit, but the new condition is going to be more difficult to express and we are going to obtain no substantial advantage).

In any case, what we are going to do now requires no special condition on $\Vdash$ and can be dealt with in the greater generality.

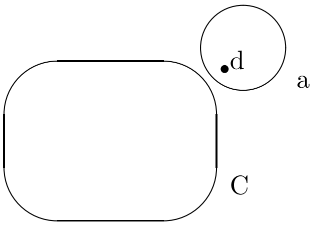
The main idea is now that we can use the elements of the set of the properties in order to collect different documents or also to separate some documents from other ones.

The two main notions in topology are the notions of closed and open subset.

A closed subset is a subset which contains all the points which are not "far enough" from it. One can think of a closed subset C like of a magnet: given the magnet, it is going to grow by attracting all the things which cannot be safely separated from it. This condition explains the chosen name. Let us suppose now that a document d is safely separated if a property can be found which is satisfied by d but by no point in the closed subset C, then the document d is not in C. We can express this condition formally by writing

$$d \notin C \text{ if and only if there exists } a \in P \text{ such that } d \Vdash a \text{ and } \neg(a \between C)$$

where $a \between C$ means that there is at least one document in C which enjoys the property a.



This condition can be expressed also in a positive way if we say that a document belongs to a closed subset C if it cannot be separated by the closed subset; formally

$$d \in C \text{ if and only if for all } a \in P, \text{ if } d \Vdash a \text{ then } a \between C$$

So, provided we know how to construct the smallest closed subset C which contains some wanted documents, we can use C in order to collect all the documents which are not "far enough" (or better "too far") from the prescribed ones, that is, all the documents such that there is no property that they satisfy and that is satisfied by no document in C.

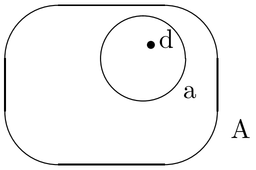
On the other hand, we can be interested in all the documents which satisfy some given properties $a_1, \ldots, a_n \in P$. If we want all the documents which satisfy *all* of the required properties, and the second condition that we proposed on $\Vdash$ in order to have a topology holds, then we can find directly the wanted subset since it is

$$\{d \in D \mid d \Vdash a_1 \ \mathsf{and} \ldots \mathsf{and} \ a_n\}$$

But we can also be interested to all the documents which satisfy *at least one* of the required properties. Then the topological notion of open subset is what we are looking for. An open subset A is a subset which let leave all the points which are not "near enough", that is

$$d \in A \text{ if and only if there exists } a \in P \text{ such that } d \Vdash a \text{ and } a \subseteq A$$

where $a \subseteq A$ means that all the points which satisfy the property a are points in A. The name "open subset" was chosen to suggest this situation.



So, if A is an open subset, it can be obtained like the union of all the points which satisfy at least one of the properties a such that $a \subseteq A$.

Open and closed subsets are the fundamental ingredients of every topology and by using them we can specify the search criteria that our search engine has to satisfy. For instance, let us suppose that we want all the documents which are not "too far" from those those which satisfy all of the properties a_1, $\ldots$, a_n; then the subset that we are looking for is the smallest closed subset which contains the open subset $A \equiv \{d \in D \mid d \Vdash a_1 \ \mathsf{and} \ldots \mathsf{and} \ a_n\}$, namely the closure $\mathsf{Cl}(A)$ of the subset A. $\mathsf{Cl}(A)$ can be defined by putting

$$\mathsf{Cl}(A) \equiv \{d \in D \mid (\forall a \in P) \ d \Vdash a \rightarrow a \between A\}$$

Indeed it is possible to show that $\mathsf{Cl}(A)$ is a closed subset which contains A and any closed subset which contains A contains also $\mathsf{Cl}(A)$. Here is the proof. Let us first prove that for any subset A of points and for any property a,

$$a \between A \text{ if and only if } a \between \mathsf{Cl}(A)$$

To prove the left to right implication, first note that $A \subseteq \mathsf{Cl}(A)$. In fact, if $d \in A$, $b \in P$ and $d \Vdash b$ then $(\exists y \in D)\ y \Vdash b\ \&\ y \in A$, hence $(\forall p \in P)\ d \Vdash b \to b \between A$, that is $d \in \mathsf{Cl}(A)$. Then obviously $a \between A$ yields $a \between \mathsf{Cl}(A)$. To prove the other implication, let us suppose that $a \between \mathsf{Cl}(A)$, that is $(\exists y \in D)\ y \Vdash a\ \&\ y \in \mathsf{Cl}(A)$, then $(\exists y \in D)\ y \Vdash a\ \&\ (\forall b \in P)\ y \Vdash b \to b \between A$; hence by a logic step, namely existential elimination, we obtain that $a \between A$.

It is now immediate to see that $\mathsf{Cl}(A)$ is a closed subset; in fact

$$\begin{array}{lll} d \in \mathsf{Cl}(A) & \text{if and only if} & (\forall b \in P)\ d \Vdash b \to b \between A \\ & \text{if and only if} & (\forall b \in P)\ d \Vdash b \to b \between \mathsf{Cl}(A) \end{array}$$

We can now show that $\mathsf{Cl}(A)$ is the smallest closed subset that contains A. We have already proved that $A \subseteq \mathsf{Cl}(A)$. So, let us suppose that C is any closed subset that contains A. Then $\mathsf{Cl}(C) = C$. In fact, we already proved that for any subset C, $C \subseteq \mathsf{Cl}(C)$ holds; moreover $d \in \mathsf{Cl}(C)$ if and only if $(\forall b \in P)\ d \Vdash b \to b \between C$ if and only if $d \in C$ since C is closed. Now, if $A \subseteq C$ then $\mathsf{Cl}(A) \subseteq \mathsf{Cl}(C)$ because $d \in \mathsf{Cl}(A)$ if and only if $(\forall b \in P)\ d \Vdash b \to b \between A$ which yields $(\forall b \in P)\ d \Vdash b \to b \between C$ that is $d \in \mathsf{Cl}(C)$. Hence if $A \subseteq C$ and C is closed then $\mathsf{Cl}(A) \subseteq C$.

2 Formal topology

It is interesting to note that in all what we did till now we never used negation. This seems to be useful when we are required to perform a search because a negation usually requires to search all the *space*. On the other hand, we used implication in the definition of the closure of a set, and closure also require a universal quantification on the set of the properties and an existential quantification on the set of the documents. Since such sets can be huge (in particular the set of documents!) it can be convenient to work only with the set of the properties. To this aim, we need to define two operators which let us specify any open and closed subset by using only subsets of properties. We will then be able to work only with subsets of the set of properties, which are in general easier to deal with, and to move to the subsets of the set of documents only *after* having performed all the required computations.

The task of specifying an open subset A by using only the properties is not too difficult since any open subset is just the union of all the subsets whose elements satisfy the properties "contained" in the open subset itself, namely

$$A = \bigcup_{a \subseteq A} \mathsf{ext}(a)$$

where $\mathsf{ext}(a) \equiv \{d \in D \mid d \Vdash a\}$. So, in order to specify A we can simply use the subset U_A built with all the properties which are contained in A, that is

$$U_A \equiv \{a \in P \mid a \subseteq A\}$$

It is now possible to show that given any subset U of properties

$$\mathsf{Ext}(U) \equiv \bigcup_{a \in U} \mathsf{ext}(a)$$

is an open set and that, given any open set A, the subset U_A is a subset of properties such that $\mathsf{Ext}(U_A) = A$ [SV00].

Something similar can be done also for closed subsets, but it is first necessary to discover how to determine a closed subset by using only the properties. A key observation is that a closed subset C is determined by the subset of the properties which *meet* it, that is, by the subset

$$F_C \equiv \{a \in P \mid a \between C\}$$

In fact, it is possible to show that two closed subsets C and D are equal if and only if $F_C = F_D$ [SV00]. The next step is to find a link between subsets of properties and closed subsets, namely the inverse of the definition of the map F_C which associates a subset of properties with a closed subset. Let us put

$$\mathsf{Rest}(F) \equiv \{d \in D \mid (\forall a \in P)\ d \Vdash a \rightarrow a \in F\}$$

and we obtain that, for any subset F of properties, the subset $\mathsf{Rest}(F)$ is a closed subset and that, given any closed subset C, the subset F_C is a subset of properties such that $\mathsf{Rest}(F_C) = C$ [SV00].

In order to be able to work directly with the properties, it is convenient to introduce some abbreviations. We will write

$$a \lhd U \text{ to mean } a \subseteq \mathsf{Ext}(U)$$

and

$$a \ltimes\!\!\rtimes F \text{ to mean } a \between \mathsf{Rest}(F)$$

Then some useful conditions on $\lhd$ and $\ltimes\!\!\rtimes$ can be found. We will show here some of them.

$$\textit{(reflexivity)} \qquad \frac{a \in U}{a \lhd U}$$

$$\textit{($\lhd$-transitivity)} \qquad \frac{a \lhd U \qquad U \lhd V}{a \lhd V}$$

$$\textit{(}\mathsf{and}\textit{-right)} \qquad \frac{a \lhd U \qquad a \lhd V}{a \lhd U \ \mathsf{and}\ V}$$

$$\textit{(Anti-reflexivity)} \qquad \frac{a \ltimes\!\!\rtimes F}{a \in F}$$

$$\textit{($\ltimes\!\!\rtimes$-transitivity)} \qquad \frac{a \ltimes\!\!\rtimes F \qquad b \in G\ [b \ltimes\!\!\rtimes F]}{a \ltimes\!\!\rtimes G}$$

$$\textit{(Compatibility)} \qquad \frac{a \ltimes\!\!\rtimes F \quad a \lhd U}{U \ltimes\!\!\rtimes F}$$

where $U \lhd V$ is a shorthand for the proposition $(\forall u \in U)\ u \lhd V$, $U\ \mathsf{and}\ V$ is a shorthand for the subset $\{u\ \mathsf{and}\ v \in P \mid u \in U \text{ and } v \in V\}$ and $U \bowtie F$ is a shorthand for the proposition $(\exists u \in U)\ u \bowtie F$.

This list is not complete but it can be useful in any case to begin to work with $\lhd$ and $\bowtie$.

We can now show one small example of the use of $\lhd$ and $\bowtie$. We observed that in order to find all the documents which are not too far from some specified ones we have to be able to construct the closure of an open subset. So, let us suppose that the open subset is determined by the subset of properties U. Then the subset of properties which determines, by means of the operator Rest, the closure of $\mathsf{Ext}(U)$ is $F \equiv \{a \in P \mid (\{a\}\ \mathsf{and}\ U) \bowtie P\}$. In fact

$$\begin{array}{lll}
d \in \mathsf{Rest}(F) & \text{if and only if} & (\forall a \in P)\ d \Vdash a \rightarrow a \in F \\
 & \text{if and only if} & (\forall a \in P)\ d \Vdash a \rightarrow (\{a\}\ \mathsf{and}\ U) \bowtie P \\
 & \text{if and only if} & (\forall a \in P)\ d \Vdash a \rightarrow (\exists y \in D)\ y \Vdash a \ \&\ y \in \mathsf{Ext}(U) \\
 & \text{if and only if} & d \in \mathsf{Cl}(\mathsf{Ext}(U))
\end{array}$$

3 Conclusion

It is clear that what was presented till now is just the beginning of a new research topic. At least two open problems are immediately evident: to find all the valid conditions on $\lhd$ and $\bowtie$ since they are going to be necessary in order to be able to work only with the elements of the set of properties and to find some feasible algorithm to discover when $a \lhd U$ and $a \bowtie F$ hold since this is an essential condition to be able to discover the open and the closed subsets.

References

[SV00] Valentini, S. The problem of completeness of formal topologies with a binary positivity predicate and their inductive generation, to appear.