# CS 6630 Project Process Book
# Visualizing SIGGRAPH Publications

Kui Wu and Duong Hoang
u0931640 and u0930062
u0931640@utah.edu and u0933062@utah.edu

November 15, 2015

## 1 Introduction and background

When surveying a new field, researchers often want an overview look at the important papers in the field, how they are related, and how popular trends in topics or techniques have been evolving over the years. Traditional search engines such as Google Scholar and digital libraries such as the ACM's can alleviate the task, but only when the researcher has a good idea of what he should look for. Even then, the information obtained from those tools at any given moment is not only too specific but also presented in a non-visual form, making it difficult for the user to see a bigger picture.

Being both interested in doing research in computer graphics, in this project we aim to develop a tool to visualize connections between papers published over the years in SIGGRAPH, the most important conference in computer graphics. The tool should show in one place: the citation relationships among SIGGRAPH papers, important papers in each sub-field, prolific authors and their collaboration patterns, popular topics and methods in recent years, and active research institutions in each sub-field. It is our hope that such a tool can be particularly helpful to someone who wants to survey a field and summarize major results. It could also help new researchers in finding not only interesting problems to work on but also pointers to relevant publications for reference. Lastly, prospective graduate students can use information provided by the tool to make more informed decision in their application process.

When searching for project ideas, we stumbled upon Autodesk's Citeology and this work gives us motivations to propose the current project.

## 2 Project Objectives

Our project aims to answer the following questions:

- What papers are influential in the field, in terms of citation count?
- Given a paper of interest, what papers does it cite? What are the papers citing it?
- What topics and techniques are popular, and in which time periods?
- Given a set of keywords, what are the most relevant papers/authors/institutions to these keywords?
- Who does each author collaborate with the most?
- What institutions are more active in a given field, in terms of publication count?

Answering the above questions will put a make a researcher more informed about what papers/authors/topics/techniques to pay more attention to, thereby saving time in the beginning of a survey task. It can also reveal interesting patterns and trends in the past that can help him or her better decide on future research directions for a particular topic of interest.

# 3 Data

Our data comes from multiple sources.

- Paper texts are downloaded from the ACM Digital Library. From these we extract the title, authors and their affiliations, references, and important keywords.
- BibTeX information for all papers is downloaded from the Digital Bibliographic Library Browser.
- Citation count information is queried from Google Scholar.

# 4 Data Processing

We expect our raw data to require a substantial cleanup process. So far we have collected all the SIGGRAPH papers's pdf files from 2002 to 2015, totaling 17GB of data. These pdf files will be converted to text format, from which we will extract from each paper its title, authors and their affiliations, references and keywords. The extraction process is done using a combination of homemade Python scripts and opensource tools such as ParsCit.

Meaningful keyword extraction is the most difficult step of the extraction process. Most papers provide a set of "keywords" after the abstract but in our experience, these keywords are not meaningful enough: in any given year, most keywords are used by only one paper. We are looking into using an auto-summarizing tools for this step. Because of the uncertainty of keyword extraction, all features that require keywords will be optional.

We have done cursory testing of our tools and they seemed to be able to extract the required information with high accuracy, thanks to the uniformity in format of the papers. There is still some chance that noise will show up in the extracted data, in which case some post-process cleaning has to be done. We will finally need to remove all non-SIGGRAPH papers from the list of references to make our data more self-contained and easier to work with.

Since our data is highly relational, we plan to store it in an SQL database and query the data with Javascript. At the moment sql.js and node-sqlite3 seem to be promising candidates. Tentatively, there will be different tables, one for each of the following: papers, authors, institutions, keywords. Each entry in a table will have necessary fields, for example a paper will contain links to its authors and keywords, an author entry will have links back to his or her papers, and (optionally) links to the institution where he/she worked when a paper was published.

An alternative design is to flatten all the SQL tables into JSON files and load those instead. The SQL approach allows us to not having to load everything into memory at once, perhaps at the cost of more processing time each time users change their query/filter of the data. It is not yet clear to us how big the data will be, and which one of the two approaches will work better.

# 5 Visualization Design

## 5.1 Proposed Design

Our design will consist of four different views (Fig. 5):

**Paper view (the main view)**. Here we show all the papers, grouped by year. Each year occupies one column. This view will be wide enough to show about ten years but will be scrollable horizontally to move the focus to a different period of time. Each paper's title will appear on one row and be click-able. When a paper is clicked on, the papers that this paper cites and the ones that cite it will be highlighted (Fig. 6). Moreover, when a paper is moused over, a pop-up will show the paper's full title, author lists, keywords, and DOI link (Fig. 7). A bar will appear right under each paper's title, the length of which is proportional to the paper's number of citations (this is not yet shown in our design sketches). This help the user identify influential papers at a glance.

This whole paper view can be sorted either alphabetically or by citation count. The papers can also be filtered by citation count, to hide papers with lesser impacts. We provide two checkboxes and a slider for these purposes.

**Institution view**. Here we show a map of all the institutions that have published papers to SIGGRAPH. They will appear as circles on a projected world map. The bigger the circle is, the more publications that institution has in the selected time period. Mousing over an institution will show its name and address. We chose a map for this view because for an institution, the geographical information can be important, for example, to a prospective

graduate student looking for a school to apply to. Displaying a large amount of items using circles is also space-conserving.

**Author view**. In this view we show all paper authors and their collaboration relationships as a node-link diagram. Each author is a node, and two authors are linked if they have written a paper together. Bigger nodes represent more prolific authors, in terms of some metrics such as the H-index. Mousing over an author will show his/her name and affiliations. The reason we chose a node-link diagram for this view is because it highlights quite nicely the clusters between subsets of nodes, and the view is dynamic so it is easier to put any author at the center (for example, when his paper is being selected).

**Keyword view**. This view acts both as a view and a filter. Here we list all the keywords extracted from all the papers in alphabetical order. Keywords are selectable, and each time a keyword is clicked on, the related keywords are highlighted. This feature is particularly useful when the user clicks on a topic keyword (for example: "global illumination"), and the related keywords show common techniques used to solve problems related to that topic (for example: "path tracing", "radiosity", "photon mapping" which are common methods in graphics to achieve "global illumination"). Two keywords are related if they appear together in many papers. Moreover, selecting a set of keywords will reduce the amount of information shown in the other views, to retain only the papers/authors/institutions that are related to the selected set of keywords. This is useful because the list of papers/authors/institutions can be quite large. Fig 8 shows an initial design of this idea, where the list of papers was not really filtered by simply highlighted and re-sorted. This feature can be very interesting because a glance at the paper view can tell us the popularity of the current selected set of keywords throughout the years.

**Interaction between views**. In the previous section we described what happens to the other three views when a keyword is selected. All the elements in the first three views can be clicked on as well. When a paper is clicked on, the corresponding authors will be highlighted and brought to the center of the view, and the corresponding institutions will also be highlighted. We will also highlight the selected paper's keywords. See Fig. 9 for an illustration of this.

Similarly, an institution or an author can be clicked on, and corresponding papers and keywords will be highlighted as well (See Fig. 10).

Finally, the paper view supports the use of a brush to limit the data in other views to a particular period of time (see Fig. 11).

## 5.2 Alternative Designs

We thought about two other alternatives to represent the keyword list: a radial Reingold-Tilford Tree (12), and a zoomable multi-layer ring (13). These ideas are inspired by designs we found on Mike Bostock's website. We decided that these designs make it harder for users to read the keywords, and our keyword list is flat rather than a hierarchical and thus would be unsuitable for these hierarchical designs.

For the paper view, we thought about using lines to show citation relationship between

papers. However these lines would block underlying papers, and feel redundant because we can encode the citation relationship no less effectively using less "ink" in our proposed design.

# 6 Features

## 6.1 Must-Have Features

The paper, institution, and author views are must-have features. All the proposed interactions between these three views are essential and must be implemented.

## 6.2 Optional Features

Because it is considerably harder to extract meaningful keywords out of a paper, we will implement the keyword view and all related interactions last. In the event that it becomes too hard to do, this view will be left out of the design.

The details pop-ups that appear when mousing over each paper will also be an optional feature as they do not affect the core usability and usefulness of the project.

# 7 Project Schedule

- Week 10: Process data
  - Extract data, such as citation counts and affiliations, from papers and Google Scholar.
  - Create an SQL database based on the extracted data.
- Week 11: Create a basic framework
  - Create "Author" view, "Paper" view, and "Institution" view.
- Week 12: Project Milestone due
  - Create interactions between views. Selecting a element in one view will update the other views accordingly.
- Week 13: Optimization
  - Optimize interaction speed if necessary
  - Improve the aesthetic and visual coherency of the design
- Week 14: Extract keywords
  - Extract keywords and insert them into the database

  - Create "Keyword" view and link it to the others views

- Week 15: Final Project due
  - Improve the project based on the instructor's feedbacks

# 8 Project Progress - Milestone 1

In this section we detail the progress of the project at the point of the milestone (13 November 2015). At this stage, we have devised and tested a working plan for data acquisition and processing, and have almost done executing this plan to obtain all the necessary data. We have also implemented a working prototype for the paper view where we show all the papers and the user can select a year and a paper to see its referenced papers and papers citing it.

## 8.1 Data Acquisition and Processing

Our data acquisition and processing pipeline is as follows:

- We start by downloading all SIGGRAPH publications (in PDF format) since 2002 from the ACM Digital Library.
- Download BibTeX files (in XML format) for all SIGGRAPH and TOG papers from the Digital Bibliographic Library Browser
- Use a PDF processing tool to perform OCR and convert all the PDF files to TXT.
- Based on the TXT files, use ParsCit to extract the title and abstract from each paper.
- Use a homegrown tool (written in Python) to extract the list of references from each paper.
- Use another homegrown tool (in Python) to query Google Scholar to get the number of citations for all papers. This step is needed because we want to account for citations from non-SIGGRAPH papers as well.
- To extract the keywords, we use a Web service called Alchemy API (from IBM). The keywords returned by this tool will be combined with the keywords extracted directly from each paper's Keywords section.
- Finally we use a Python script to combine all the extracted information into a data structure in memory, from which we write several JSON files ready to be used by our Vis scripts.

At the moment, we have finished extracting all the essential information (title, author, year, references, citation count, DOI link) from our paper database, and produced working JSON files that have been used in our prototype. Below is an excerpt from our main JSON file:


```
"2002": [
{
    "authors": [
    "Jeffrey Smith",
    "Jessica K. Hodgins",
    "Irving Oppenheim",
    "Andrew P. Witkin"
    ],
    "cited_by": [
    1349,
    1384,
    1543,
    1752,
    1704,
    1846
    ],
    "id": 69,
    "link": "http://doi.acm.org/10.1145/566654.566580",
    "references": [],
    "title": "Creating models of truss structures with optimization",
    "year": "2002"
},
{
    "authors": [
    "Timothy J. Purcell",
    "Ian Buck",
    "William R. Mark",
    "Pat Hanrahan"
    ],
    "cited_by": [
    163,
    124,
    134,
    93,
    120,
    190
    ],
    "id": 63,
```

```
    "link": "http://doi.acm.org/10.1145/566654.566640",
    "references": [],
    "title": "Ray tracing on programmable graphics hardware",
    "year": "2002"
},
```


This database will be updated once the keyword extraction step finishes, in at most two days, at which point there will be no data gathering tasks left. Compared to what was written in the proposal, it turned out that extracting the affiliation information for the authors from the paper texts is too difficult, hence we have decided to drop the Institution view from the project.

## 8.2 Paper View Revolution

In this section, we will give some details abotu how our paper view changed during implementation.

The main issue we met is the amount of papers in the each year is quite large, which is about 150 papers per year. As a result, there is no way to arrange the entire paper view, 14 years and 100 papers per year, to one view with proper font size. The initial version of our implemention is shown as Figure 1.

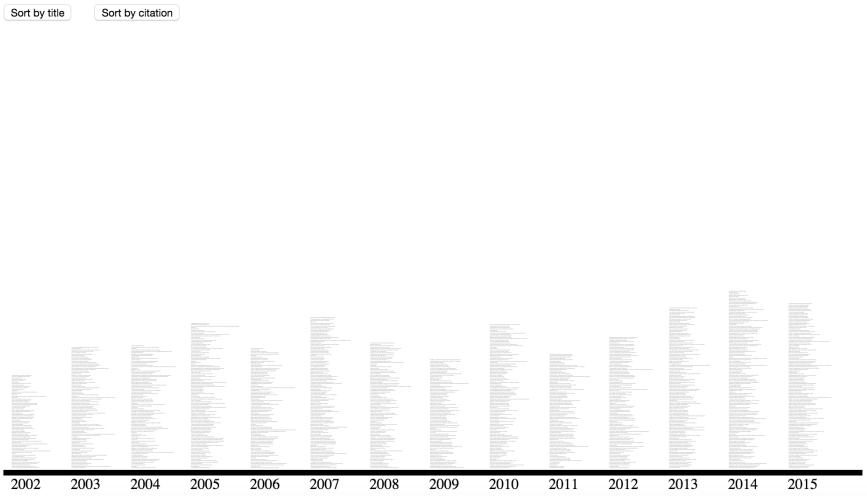


Figure 1: The initial version of our paper view implementation

There are two options to solve this issue. The first one is using zoom in and zoom out to let users explore paper views. We attempt to avoid that method because users may lose the entire view on their mind due to the change blindness. Hence, we added fisheye distortion to enable users to only zoom in texts in a small region as shown in Figure 2. However, the fisheye plugin we used is not suitable for zooming in texts. Also, because of the high density of texts, even if we rearrange texts to make it look better, it is still very messy for users to view.

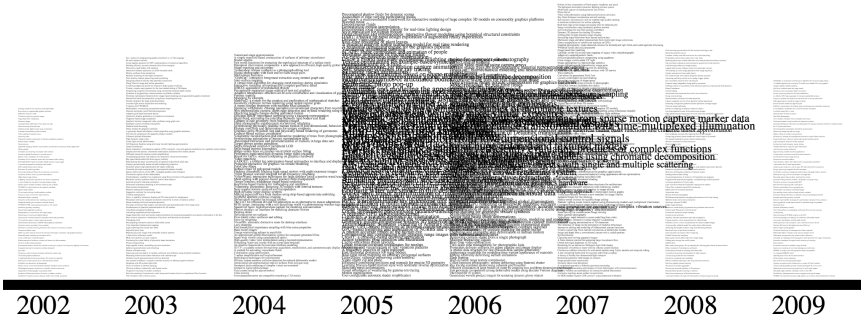


Figure 2: Paper view with fish eye

After that, we decided to add a side bar for the selected year papers as shown in Fgire 3. When users select one year, the side bar will show all papers in that year and allow users to scroll. We believe this should be the best solution for our issue without zoom in and out. Users can get details as well as overall view at the same time at least.

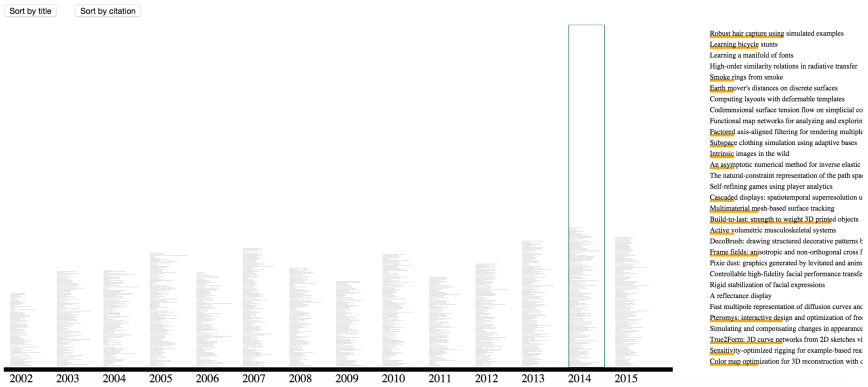


Figure 3: Paper view with side bar

However, we meet another question, which is how represent reference and citation relationship between papers in the paper view. As we mentioned before, it is too small to read and we cannot keep our original design as shown in Figure 5. Fortunately, at the worst case, the total amount of reference and citation by one paper is 52. We determined to enlarge the font size of those papers and arrange them by stair shape as shown in Figure 4. We use red and blue to highlight references and cited by papers for now.

Besides that, we also implemented functions, like "sort title by citation count" and "sort title by letter". More functionalities will be added before final submission.

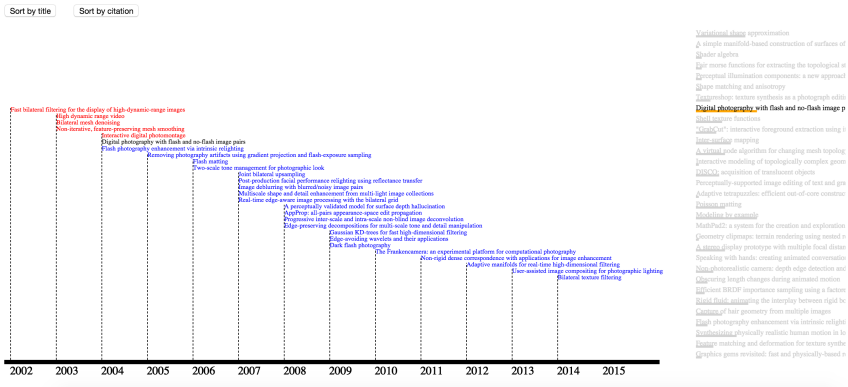


Figure 4: Paper view with reference and citation for one selected paper

During implementation, we also realized it is quite difficult to do a university maps view as we promised in the proposal because it is hard to get coordination of universities. So, we decided to get rid of it and add more functionalities to papers view instead, such as searching title. Meanwhile, we can spend more time on keyword view, which we believe is much more useful for users to explore papers.

# A Figures in Proposal


Figure 5: Overview

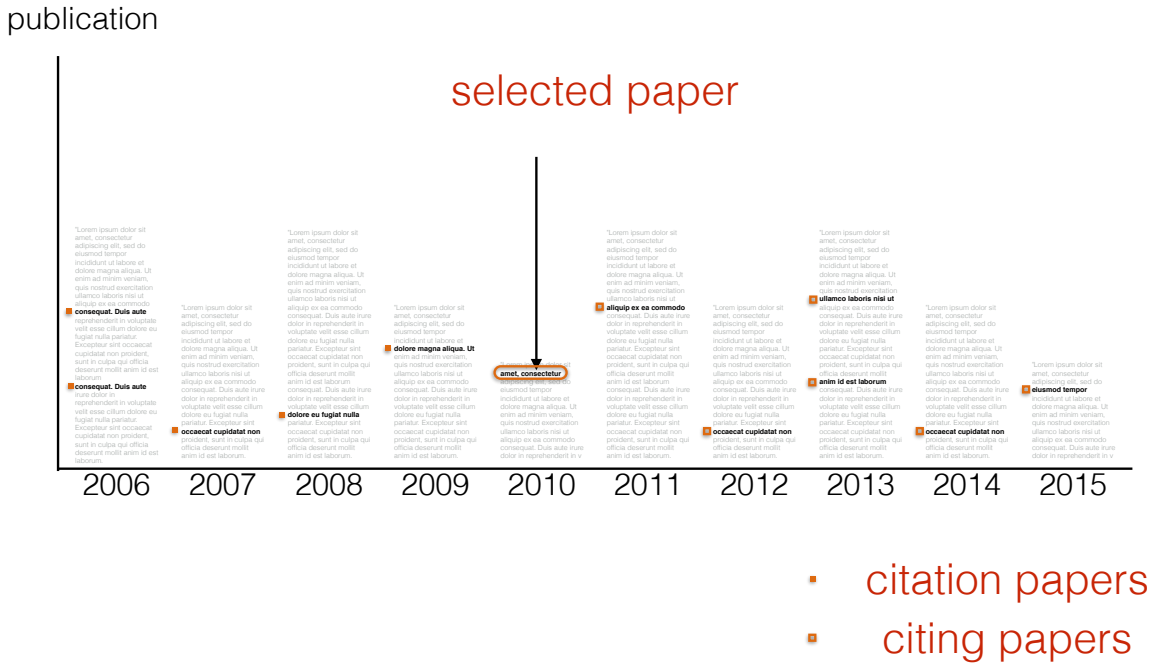


Figure 6: Paper view

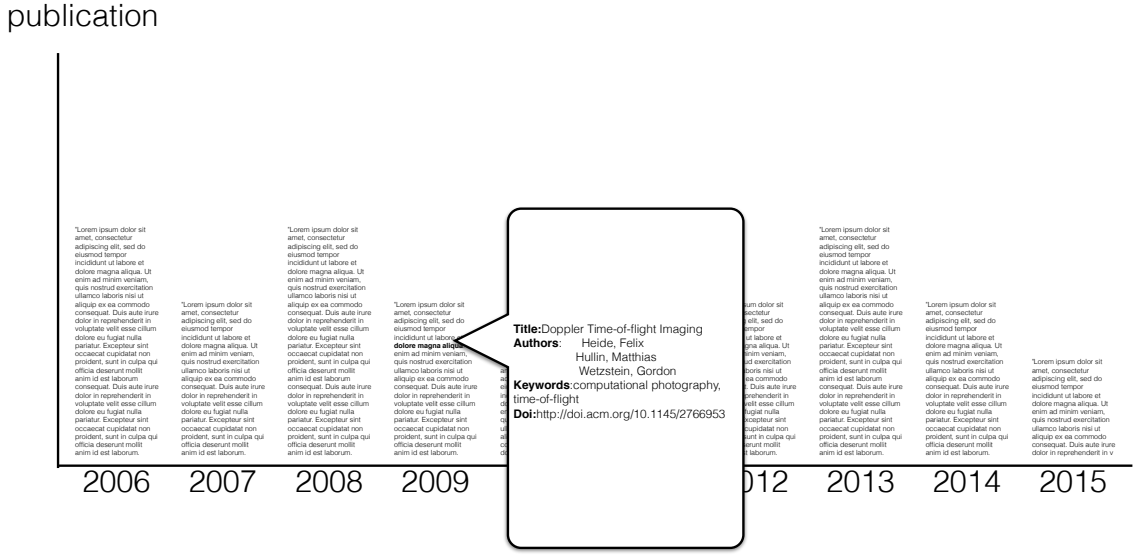


Figure 7: Paper details pop-up

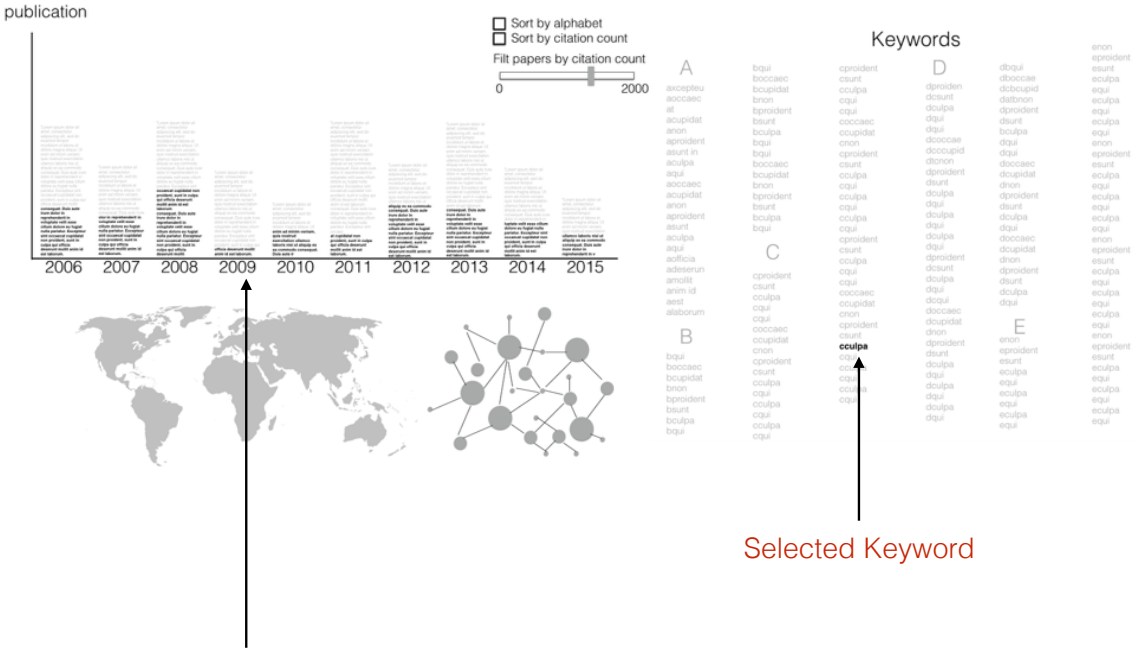


Figure 8: Keyword filtering


Figure 9: Select a paper


Figure 10: Select an author

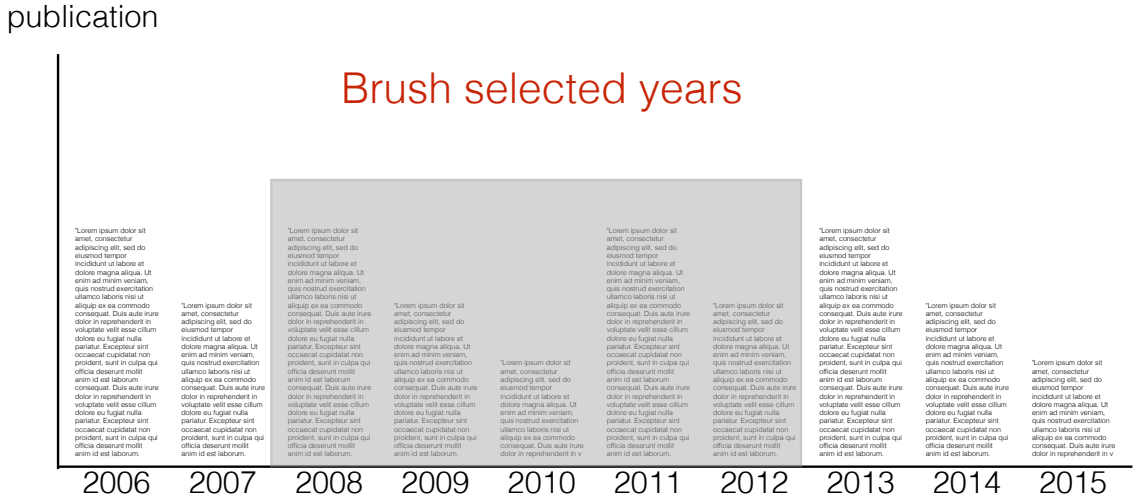


Figure 11: Brush in paper view

Keywords initial design 1:
Radial Reingold–Tilford Tree

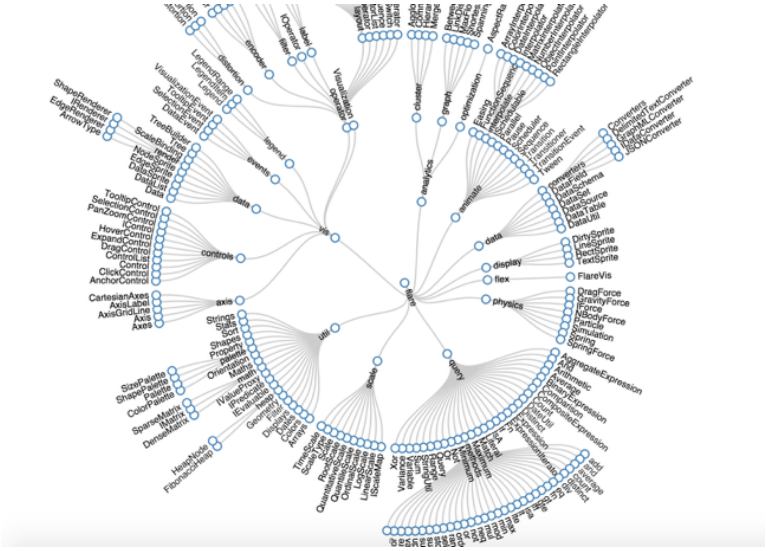


Figure 12: Radial Reingold-Tilford Tree for keywords

Keyword initial design 2:
Zoomable Sunburst

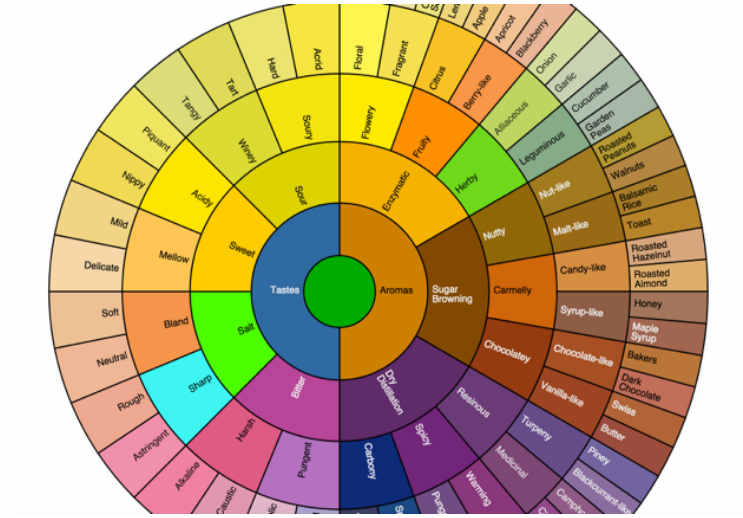


Figure 13: Zoomable rings for keywords

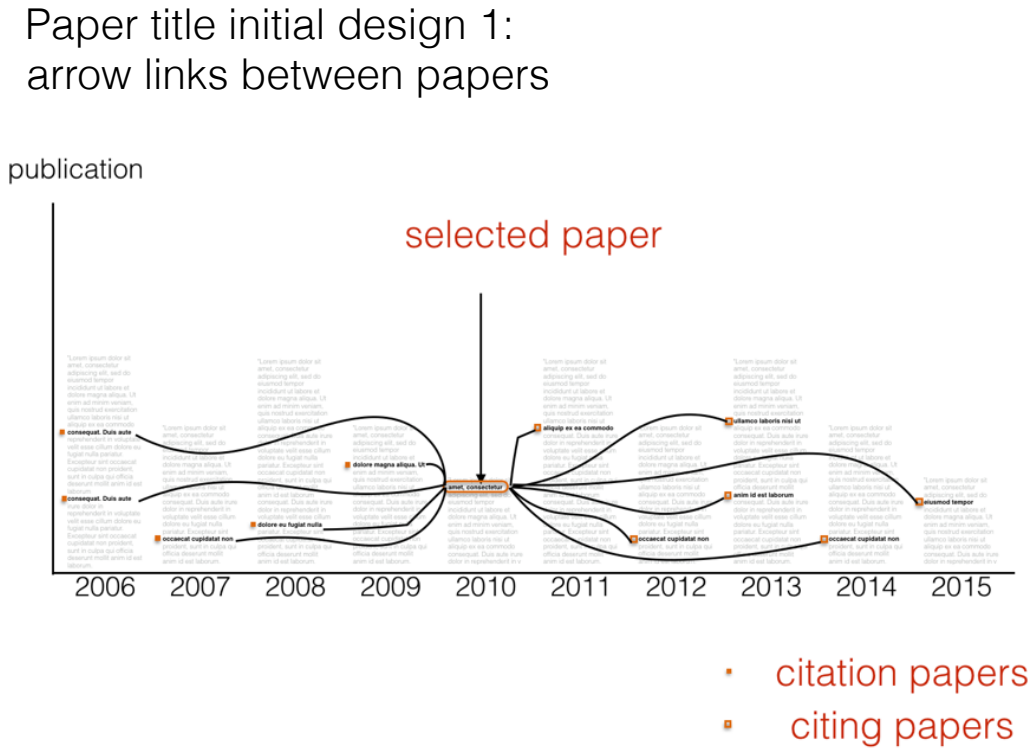


Figure 14: Arrow links between papers