Paper Trends in ION Conferences from 2007 - 2018

Adrien Perkins, *Stanford University*
Todd Walter, *Stanford University*

BIOGRAPHIES

Adrien Perkins is a Ph.D. candidate in the GPS Research Laboratory at Stanford University working under the guidance of Professor Per Enge in the Department of Aeronautics and Astronautics. He received his Bachelor of Science in Mechanical Aerospace Engineering from Rutgers University in 2013 and his Master of Science in Aeronautics and Astronautics from Stanford University in 2015.

Todd Walter is a senior research engineer at the Stanford University GPS Laboratory.

ABSTRACT

When designing a GNSS conference, one of the first steps is having individuals determine a set of tracks and sessions that are most likely to cover the topic areas of interest by those writing and submitting papers. While these decisions are made by experts in the field, it can be difficult to anticipate what topics will capture the attention of researchers and industry around the world. Furthermore, when submitting a paper, it can be hard to decide what sessions to submit a paper to when the work might lie at the intersection of several fields. As a first step to creating tools to helping the ION conference organizers decide on sessions for a conference and identifying papers that should be grouped together, this paper explores the ability to see trends in past data using commercially available natural language processing tools. Specifically, the abstract and title data for accepted papers from ION GNSS+ 2007-2018 are analyzed for different trends and patterns that can help inform future conference organization.

INTRODUCTION

The dataset used for looking at the trends presented come from just over 10 years of ION GNSS+ conference data, obtained from the ION website [1]. An important note on this dataset is that is only contains the information for the accepted papers, which limits the ability to see potentially trends that appeared in papers that were submitted but not accepted.

One of the most challenging aspects when it comes to trying to understand written data is the fact that the English language is quite variable and contains a lot of “clutter”. Therefore, to create a useable dataset, we needed to “normalize” the information contained in the titles and abstracts using natural language processing (NLP) techniques. To do this processing, we used an open source python library called SpaCy [2]. The library is maintained in such a way that it contains the most cutting edge algorithms to perform each of the NLP steps in an easy and convenient manner. The specific steps used are described in the next section.

The subsequent sections of this paper will describe the dataset itself in more detail, looking at some high level information on the papers themselves and then will dive into some of the more interesting specific trends identified in the 10 years of papers.

THE DATASET

The dataset used contains the title and the abstract information for all the accepted papers at the ION GNSS+ conferences from 2007 to 2018. To normalize the content in each of the papers (a paper being defined by a title/abstract combination), each paper was run through a lemmatization step, which transforms all words into their root (e.g. playing, plays, or played become play), and a stopword removal step, which removes words such as “the”, “a”, etc. These steps are crucial in being able to compare the words used in one paper to the words used in another paper and in removing clutter words that provide no meaningful material to the content of the paper.

From these trimmed down papers, a dictionary of “meaningful” words can be created from all the words used in all the papers from 2007 to 2018. From there each paper can be vectorized in two different ways: based on the uses of each word in the dictionary (that is, for the ith word in the dictionary, u_i will contain the number of times the ith word is used in the title and abstract) and based on the presence of each word in the dictionary. These two representations will allow us to look at words by both their raw usage and by the number of papers that are discussing a given topic.

Looking at the very first dictionary created from the ION GNSS+ data and the frequency of words used from that dictionary, depicted in Figure 1, we can hopefully start to get an idea of words that might describe the GNSS field over all. However, looking at Figure 1, one of the most used words in the dataset is the word "paper", which is not representative of the GNSS field. This can be explained by the fact that the dataset contains abstract information and abstracts typically contain phrases such as "in this paper…" or "this paper presents…". Therefore, while this is indeed an important word from an English language perspective, it is not necessarily something that is important to have in the dataset for comparing ION GNSS+ papers and provides a cautionary note on the importance of understanding the dataset and some of the limitations of these off the shelf techniques.

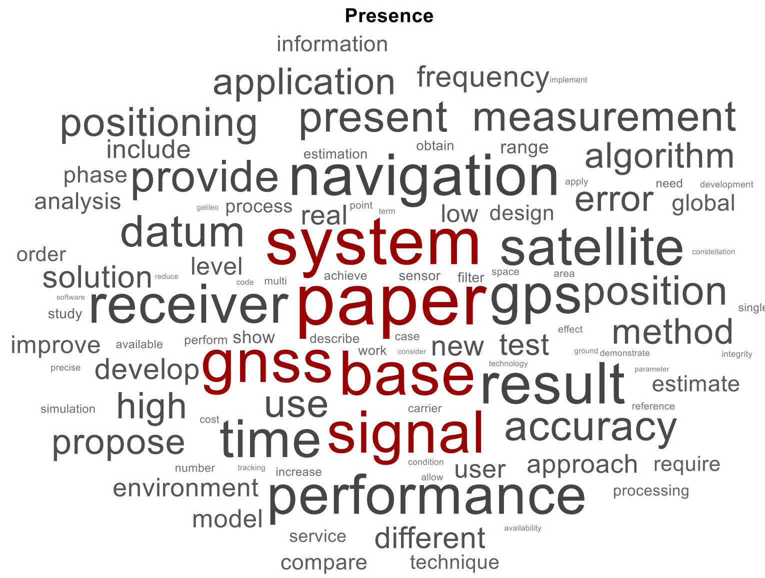

Figure 1: wordcloud of ION GNSS+ dictionary weighted by number of papers using each word

Removing the word *paper* from the dictionary, and once again looking at a word cloud weighted by frequency, results in Figure 2 and Figure 3. In this case, Figure 2 shows the word cloud weighted by the presence, or number of papers that use the given word, while Figure 3 shows the word cloud weighted by the uses, or number of instances the word appeared in the entire dataset. In the "presence" set of words, we do interestingly see the word *base* appear in many papers. This can be explained by the use of phases such as "x-based approach…" or "y-based system…" in papers. However, overall, looking at these, we see more of the words that would be expected from the GNSS field.

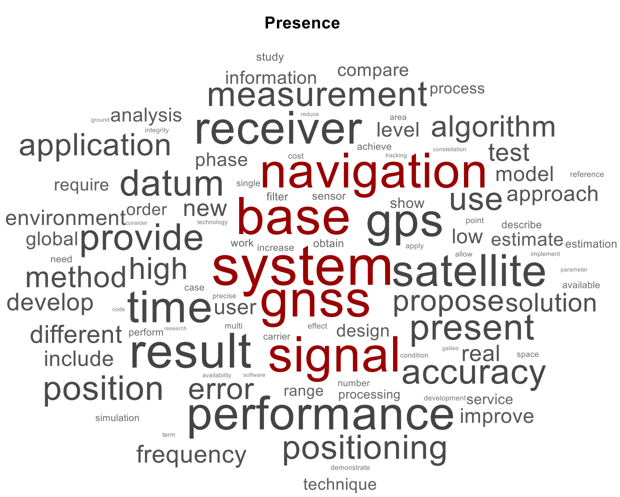

Figure 2: wordcloud of ION GNSS+ dictionary (paper removed) weighted by number of papers using each word

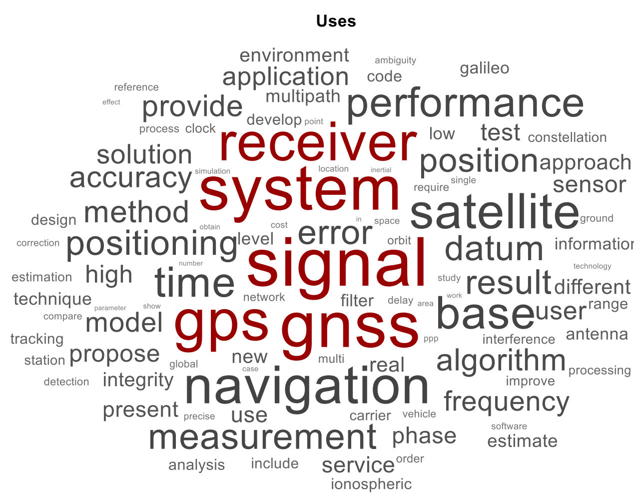

Figure 3: wordcloud of ION GNSS+ dictionary (paper removed) weighted by number of occurances of each word

The dictionary from ION GNSS+ was compared to a dictionary created from the ION ITM papers from roughly the same year period to see how similar these terms were to define the field as a whole. The word cloud from ION ITM is shown in Figure 4 and interestingly there is one important different word, the word *result*. This indicates that papers at ITM tend to be more results driven, as would be expected from the more technical conference of the set of ION conferences.

The other import note with the presence of the word *result* is the fact that we need to be a little careful when starting to remove words from the dataset (such as was done with *paper*). If we want to make the papers less similar, and highlight the non-general terms in papers, a naïve approach would be to simply remove the X most commonly used words in the dictionary (we could say those are the words that define the field), however, in this case, the word *result* is actually important to the GNSS+ conference while it may be less important to the ITM conference.

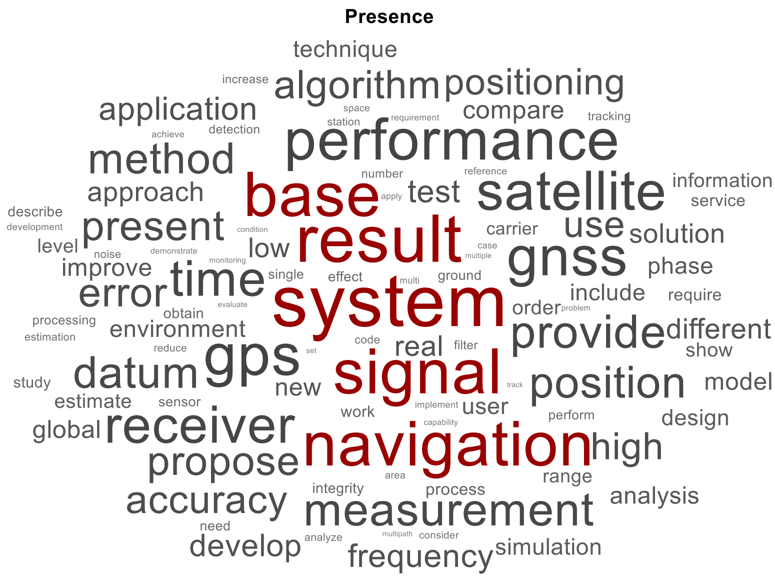

Figure 4: wordcloud of ION ITM dictionary (paper removed) weighted by number of papers using each word

Next, to just understand how verbose people are, Figure 5 and Figure 6 show the average words/title and average words/abstract for each year. In these figures, "raw" refers to the title or abstract data before any NLP was performed and "meaningful" refers to the data after being passed through the NLP steps and references the dictionary of "meaningful" words. For the post part this shows that both titles and abstracts have tended to be very similar in length over the years.

When looking at the number of words per abstract per year, one thing that immediately sticks out is the data from 2013. Looking at a histogram of the abstract lengths from a typical year (Figure 7) compared to that of 2013 (Figure 8), we can see that there is a much larger spread of abstract lengths, with a larger grouping near the 1000 word mark. It turns out that the data contained on the website for 2013 contains the abstract that was submitted by the papers while every other year contains the contents of the final version of the abstract (when the paper was submitted). It is therefore important to keep in mind that the results from 2013 may be distorted compared to the rest of the dataset.

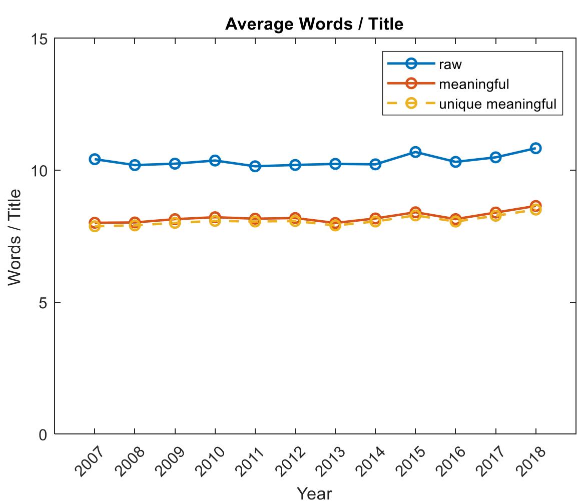

Figure 5: average words per title by year

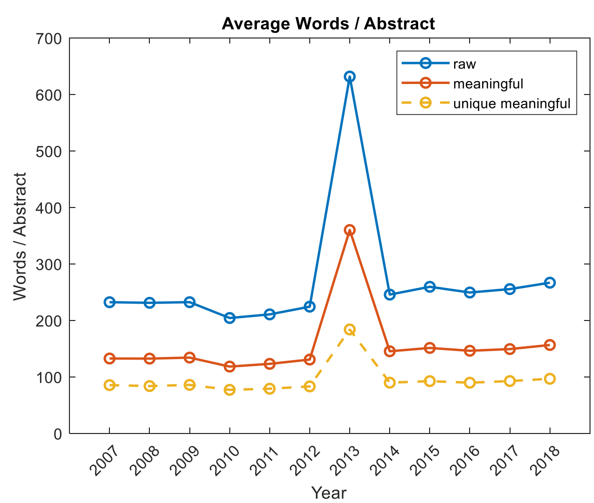

Figure 6: average words per abstract by year

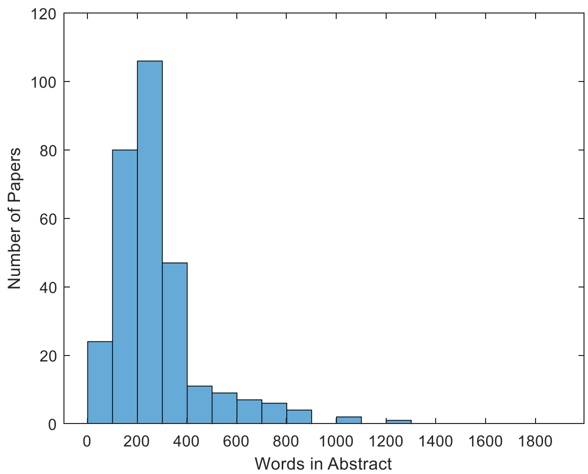

Figure 7: histogram of abstract length for 2018 (and similar to al lyears except 2013)

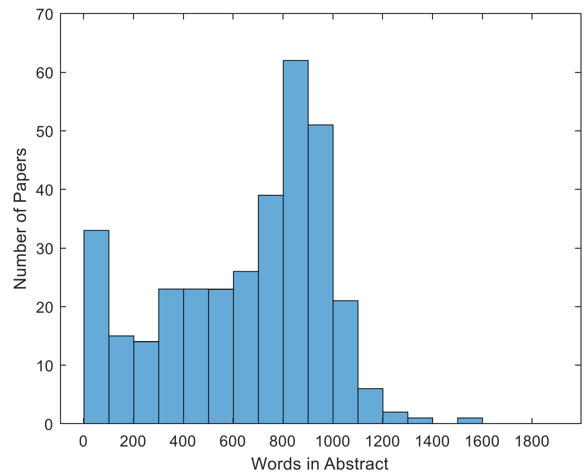

Figure 8: histogram of abstract length for 2013

TRENDS IN GNSS

With this normalized dataset of 11 years of papers, we wanted to assess the ability of specific words to provide trend information. First, we used keywords that we expected to find trends and second, we let the data define the words we should be using as keywords based on the percentage change of their appearance over the years.

Search for Trends

The first major set of keywords explored were the terms for the major constellations today: *GNSS*, *GPS*, *Galileo*, and *GLONASS* (note that all words in the dataset are lowercase, removing differences due to case). The trends for these words, both in their presence (the per paper view) and their uses (the per word view), as shown in Figure 9 and Figure 10, respectively.

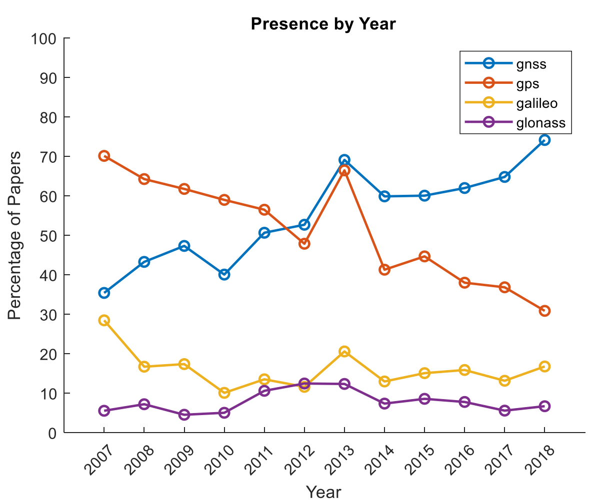

Figure 9: trends of GNSS and the major constellations by number of papers using the word

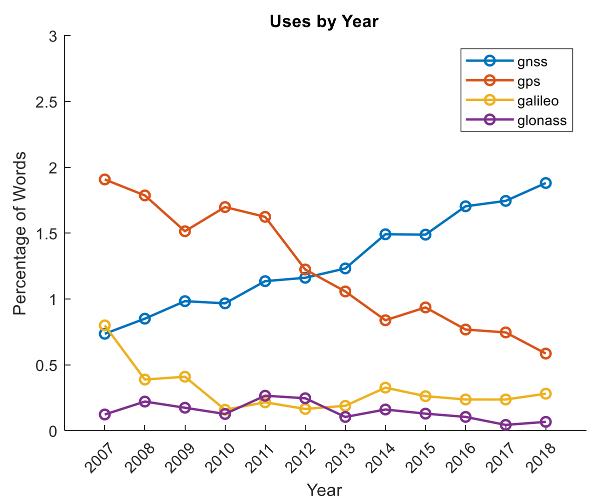

Figure 10: trends of GNSS and the major constellations by occurrences of the word

The immediate, largest trend, that can be seen is the near one-for-one trade of the word *GPS* for the word *GNSS*, indicating the widespread adoption of the word *GNSS* and the general interest in looking at not just a single constellation but rather all the constellations over the years.

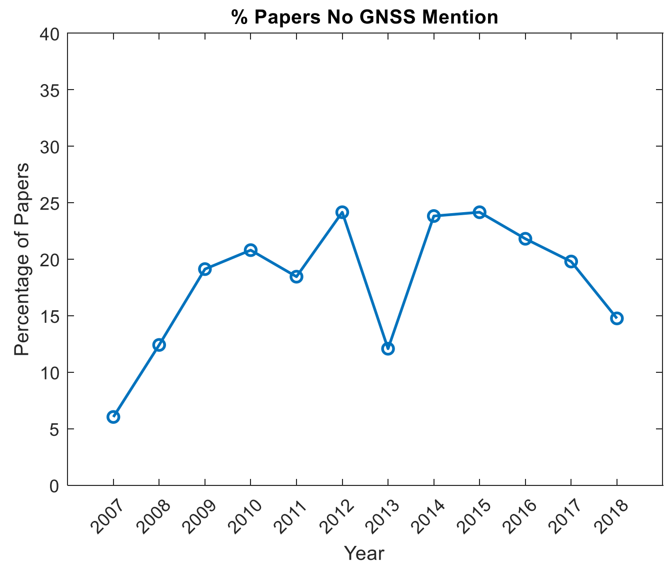

Figure 11: percentage of papers that do not reference GNSS or any major constellation

Grouping all of these terms together, and flipping the perspective a bit, we can look at the percentage of papers that do not mention *GNSS* or the name of any of the major constellations over the years in Figure 11. The first notable aspect of this plot is the fact that in 2013 there is a noticeable drop (a change of almost 15% compared to the neighboring years) in papers not talking about a constellation or *GNSS* (alternatively, 90% of the papers mentioned one of the terms while the neighboring years only had about 75% of papers mentioning the term). This once again highlights the importance of understanding the dataset as it is believed this drop is due to the fact that in 2013 the abstracts were those submitted, not the one contained in the final version of the paper. Given these were the abstracts submitted to get into the conference, it is almost expected that more papers will try and use the word GNSS in the abstract given that the conference name is ION GNSS+, while, once accepted and for the final version of the paper, the abstract tends to more closely align with the content of the paper and references to links to GNSS might be shifted to the introductory material instead of the abstract.

Looking at the collection of titles from 2018 for the papers that do not contain a reference to *GNSS* or any specific system, the papers predominantly covered topics such as autonomy, mapping, indoor navigation, and general alternate navigation methods, exemplifying the growing diversity in the ION community.

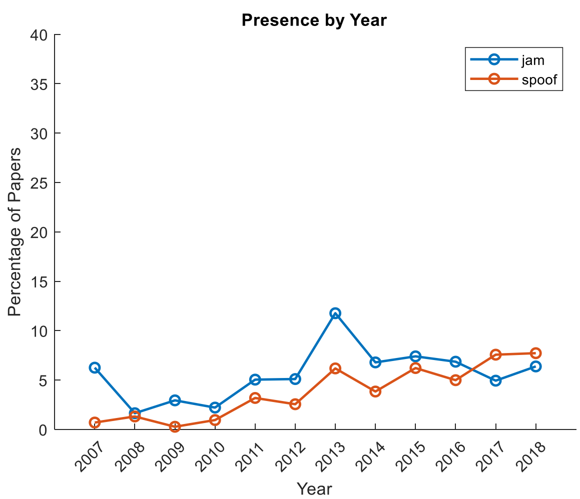

Figure 12: trends of jamming and spoofing by number of papers using the word

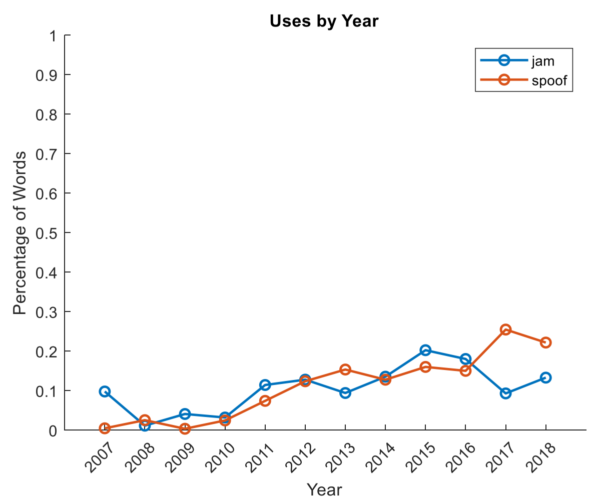

Figure 13: trends of jamming and spoofing by occurrences of the word

Other keywords searched to view expected trends are jamming (*jam*) and spoofing (*spoof*), shown in Figure 12 and Figure 13. The community interest in both jamming in spoofing has increased over the years fairly evenly.

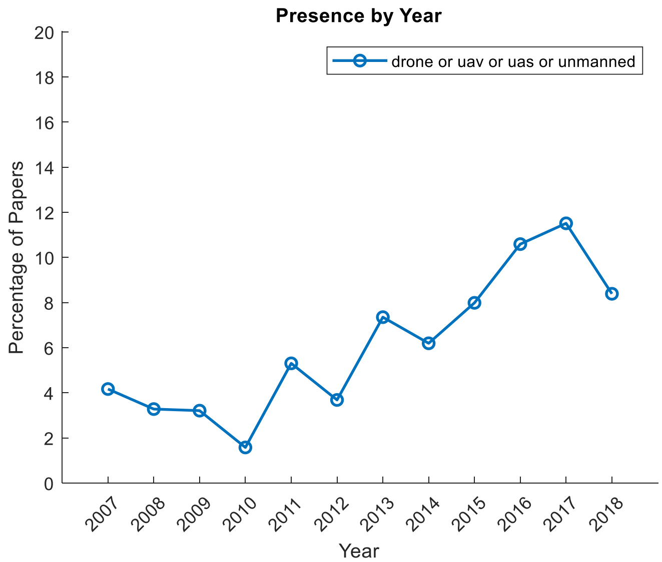

Figure 14: trends of drones by number of papers using the word

Another keyword we expected to see an increase is in drones, depicted in Figure 14. This case is another good example of some of the limitations of relying on keywords alone to identify trends, as it can be seen that to get this graph the search was not just for the word *drone*, but rather for *drone* or *uav* or *uas* or *unmanned* since all of these words are commonly used words for defining research using drones. At the moment the dictionary does not contain semantic similarity information, or the linguistic similarity, between words, requiring the user to manually make those connections in the search. As this work continues, this is one of the major areas to further develop the capabilities in order to more clearly identify these trends that go by many words.

Instead of just looking for trends, using a keyword approach can allow us to compare the popularity of similar topics within the ION community. For example, we can look at the preferred filter used by members of the community by comparing the uses of the words *ekf*, *particle*, *ukf*, and *square* for the extended Kalman filter, particle filter, unscented Kalman filter, and least squares approaches to filter. The results are shown in Figure 15 and Figure 16 with EKF being the dominant filter in the community followed by least squares. Interesting there is not necessarily a strong increase or decreasing trend for any of the filters.

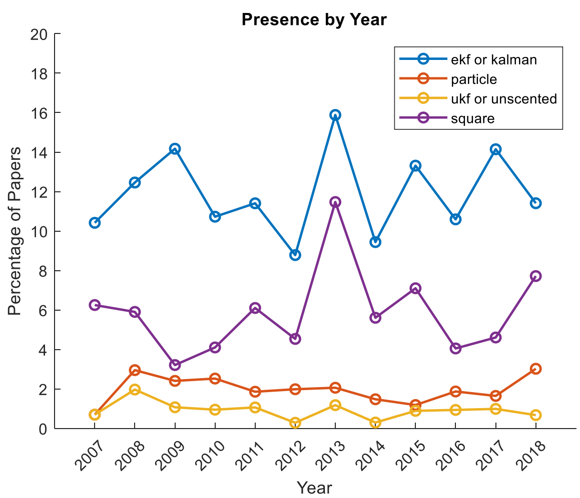

Figure 15: trends of the different filters by number of papers using the word

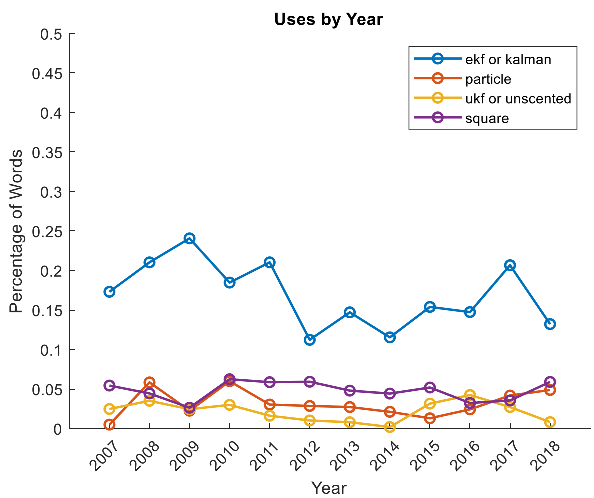

Figure 16: trends of the different filters by occurrences of the word

Percentage Change

Next we let the data identify the keywords by looking at words with the largest percent increases and decreases between 2007 and 2018. This approach can help identify general trends in the data that may not be expected.

First looking at percent increase by number of papers using the term (shown in Figure 17) we see that *smartphone* comes in at the top with a percent increase of just over 20% since 2007. Other terms that have increases are *driving*, hinting at an increased interest in cars and potentially autonomous navigation, *reduction*, *android*, and *geo*.

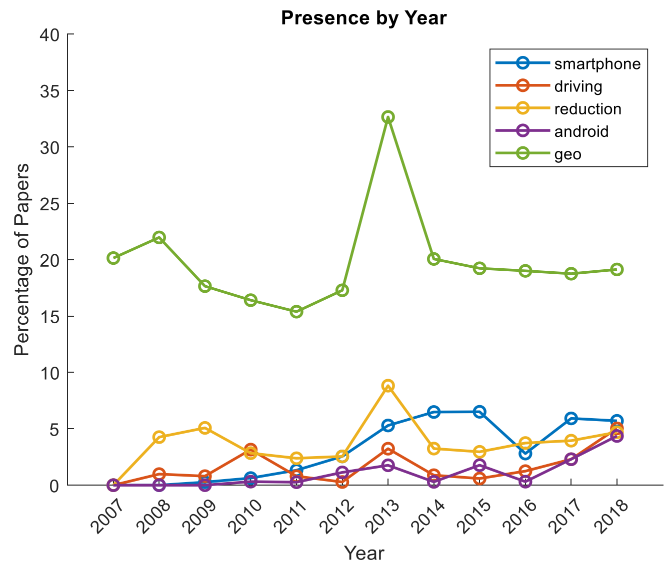

Figure 17: trends in the words with largest percent increase by number of papers using the word

Figure 18 and Figure 19 look a little more closely as the words *android* and *smartphone*. We see that the interest in smartphones has been increasing quite quickly since 2011, but it was not until after 2016 that the community showed an interest into a specific operating system. It is believed that this affect can be attributed to notable ION community member Frank Van Diggelen, who moved to Android in 2016 the subsequent launch of raw measurement on the Android platform. This highlights that member's impact on the research world can really have a profound affect in changing the direction of research interests in the ION community.

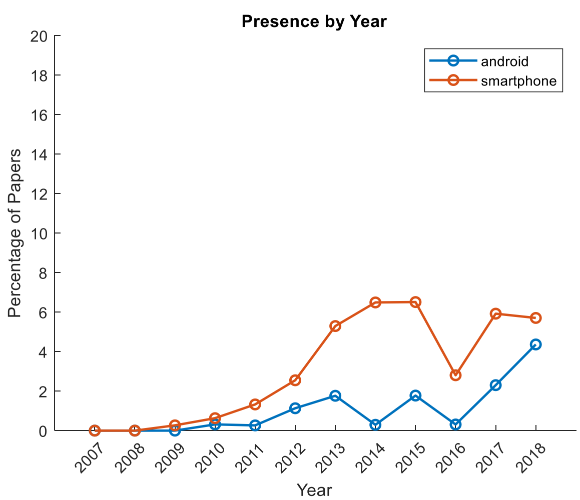

Figure 18: trends in the words smartphone and android by number of papers using the word

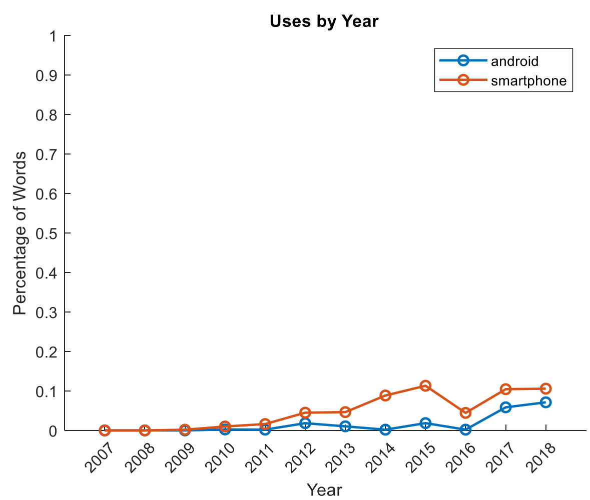

Figure 19: trends in the words smartphone and android by the occurrences of the word

On the other end, Figure 20 shows the trends for the keywords that have had the largest percent change downward; words such as *specialized*, *laas*, *preserve*, *storm*, and *giove*. Some of these terms (*specialized* and *preserve*) are harder to explain, while others are acronyms for programs that has completed their purpose (*giove*) or have fallen out of favor for other terms (*gbas* being preferred to *laas* today).

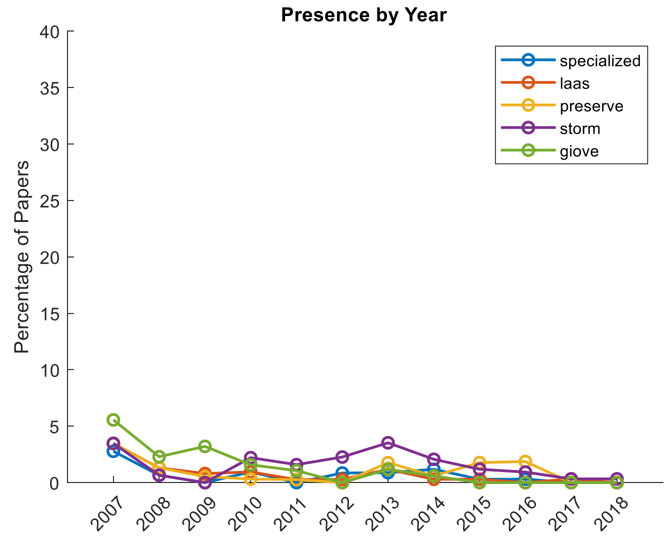

Figure 20: trends in the words with the largest percent decrease by number of papers using the word

The word *storm* was an interesting one on the list that we further explored. Figure 21 shows the trend for the word *storm* in the period from 2007 to 2018 and Figure 22 shows the solar cycle for the period from 1996 to 2018. Looking at 2009, a minimum in the number of papers using the word *storm* we see that it corresponds to a minimum in the solar cycle, while the peak use of the word in 2013 leads the solar maximum for the cycle by about a year. This shows that there are some topics of interest to the community that follow natural phenomenon that can potentially be predicted and be used to help determine sessions for conferences in the future.

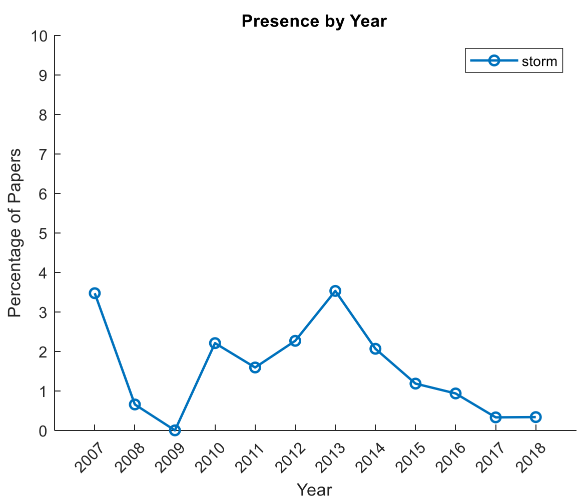

Figure 21: trend in the word storm y number of papers using the word

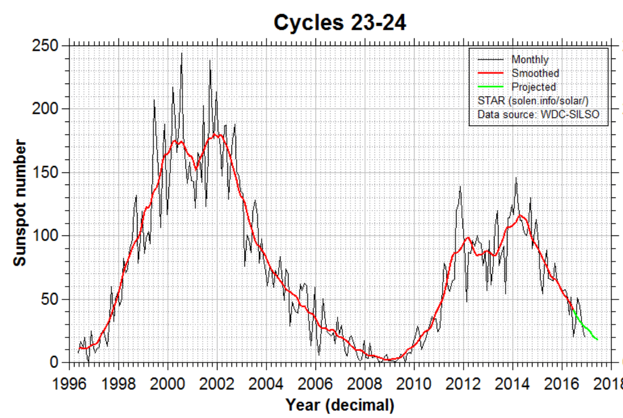

Figure 22: solar cycle for 1996 - 2018

CONCLUSION

This analysis really only scratches the surface of the possibilities of using the data of submitted papers to help conference organizers in the daunting task of creating and populating sessions with papers. The trends shown in this paper start to hint at an ability to help identify areas of interest in the community.

FUTURE WORK

The dataset used only contained the final abstract and title for accepted conference papers which naturally limit the ability of any analysis to paint a complete picture of the submission every year. One of the key first steps as this continues is to use a dataset containing the true submitted abstract and title of all the submitted papers, not just those accepted.

On the NLP side, integrating the idea of semantic similarity into the dictionary will be important and very helpful in being able to better compare two papers to each other by connecting different words together based on their underlying meanings.

Finally, this analysis so far only looks at the trends in the data and has not begun to leverage machine learning techniques to compare the contents of papers to each other. To truly help organizers populate the sessions adding the capability of comparing the contents of the papers to each other is important to find papers discussing similar content.

REFERENCES

[1] "ION." [Online]. Available: https://www.ion.org/.

[2] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," *To Appear*, 2017.