Memex II - Knowledge Creation Through Association

Dustin Lucien Smith
University of Regina
Department of Computer Science

***Abstract*— Personal knowledge bases have been a topic of discussion amongst researchers for decades. Creating a consultable and dynamic tool for knowledge creation has been a goal of many researchers throughout the last century. Through the advent of newer technologies I believe a proper interface for a knowledge base is finally possible. This report will attempt to briefly cover the history of personal knowledge bases before uncovering a probable set of 'thinking tool(s)' that have become publicly available in the recent years. Finally this report will suggest tools that may be used to further the idea of a "second brain" [5] as a thinking tool.**

I. Literary Analysis

A. History, Vannevar Bush

In 1945 Vannevar Bush wrote an article describing issues regarding the startling influx of scholarly articles and information that was being gathered. Using it as a way to describe a problem that would continue on for decades, Bush proposed that the amount a scholar may be able to recall about an article a month after reading it would be very little. He continues on referencing how "Mendel's concept of the laws of genetics" [1] was lost to the world for a generation as the writings were not accessible to those adept enough in the field to understand it. In 1945 Bush eluded to a problem that would become more widespread and prevalent as technology developed. Information is being gathered at an unwieldy rate and is being uploaded ad hoc to what we now call the world wide web. "The summation of human experience is being expanded at a prodigious rate" [1], and although our means for absorbing content has been innovated upon in the last decade, I believe it is not yet adequate to properly aggregate necessary information in highly specialized fields. Bush speaks to this problem early in his article when he writes, "There is a growing mountain of research. But there is increasing evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers - conclusions which he cannot find time to grasp, much less to remember, as they appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial." [1]

Bush carries on through his article speaking to many technical limitations of his time that have since been innovated on. Theorizing and speculating on possible solutions, Bush includes ideas for technology that were just starting to be developed post war-time, but today we take somewhat for granted. These technologies include what is modern day Optical Character Recognition, Search Engine Optimization, E-readers, and digital documents that are stored as electrical bits in solid state storage devices. Eventually Bush writes about an idea that was far ahead of his time, storing something along the lines of the Encyclopedia Britannica in such a digital format. Bush states that "Mere compression, of course, is not enough; one needs not only to make and store a record but also be able to consult it, and this aspect of the matter comes later. Even the modern great library is not generally consulted; it is nibbled at by a few." [1] This is the fundamental problem of our current storage mediums. The interface for which we consult and transact with the myriad of resources available to us is inadequate. Bush continues in his article to later rephrase the problem as being much larger than simply retrieving data, but expands his concerns to the value potential that is lost when man's acquired knowledge is lost, or is otherwise forgotten.

Bush outlines what a potential solution to his problem is in what he coined the "Memex", as "a future device for individual use, which is a sort of mechanized private file and library. ... A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory." [1] What Bush is describing throughout his article is a physical device whose sole purpose has been superseded by what we now define as a personal computer. Although his idea was limited in scope to a physical device, a digital, software based approach is equally valid.

Bush divides basic storage devices from his idea by comparing this Memex to a human brain. "The human mind ... operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain." [1], but also explains advantages that a machine may have over a brain, such as long term storage: where ideas and memories in a brain that are not frequently revisited are prone to fade, a machine's capability to 'remember' facts and figures functionally has no expiration. Finally, Bush speaks to the true value of his proposition when he offers that this artificial brain must select connections by association of idea, rather than numerically, alphabetically, or any other organizational indexing system.

B. History, Niklas Luhmann

Less than a decade after Bush talked about his post-war time ideology of scientific research, a German researcher by

the name of Niklas Luhmann was working towards a similar goal. Luhmann by the mid 1950s had begun building what he referred to in German as a "Zettelkasten", but in english translates roughly to "Note box" or "Slip box". Luhmann built a system for which he was able to categorize and store notes associatively, as opposed to linearly (as you might see in a notebook), alphabetically, or numerically. Luhmann did this through the use of a tagging system, where every note that he made had tags that could 'link' his short-form notes together by idea or frame of thoughts.

As Johannes Schmidt writes in his article "Niklas Luhmann's Card Index: The Fabrication of Serendipity", "The bulk of the collections (approximately 75,000 cards) consist of notes documenting the results of Luhmann's readings, but also his own thoughts and theoretical arguments and concepts." [8]. From what I gather, In the interest of space within his system and clarity of thought, Luhmann did not file every note he made during reading directly into the collection. Luhmann instead opted to file "what could be utilized in which way for the cards that had already been written." [7], [8]. This fundamentally changed the way Luhmann wrote his notes, as he "always [had] the question in mind of how the [documents] can be integrated into the filing system.". This was an important structural philosophy of Luhmann's slip box: that any individual idea would not be strictly filed away as being related to any one particular topic, and rather remain open for interpretation through a future lens. Furthermore, as Schmidt states: "his main concern was not to develop an idea to maximum sophistication before including the note into the collection, rather, he operated on the assumption that a decision on the usefulness of a note could only be made in relating it to the other notes." [8] In a general sense, this made any note Luhmann made particularly valuable when looking at it through a different frame of reference in the future. According to Luhmann this ideology of structuring the collection made it a "combination of disorder and order, of clustering and unpredictable combinations emerging from ad hoc selection." [8] Often the topics that Luhmann wrote about were limited to the specific subject areas that Luhmann specialized in with his research, however there was never an inherent hierarchical structure to the cards in his system. Rather it could be viewed metaphorically as what Schmidt translates as being a "web-like system" [8].

The webbed infrastructure that is created from ad hoc selection isn't without rigid rules, in fact Schmidt identifies the following 'special characteristics' of the Zettelkasten as being prerequisites for functioning: "A specific system of organization and method of card integration with specific rules of numbering, an internal system of linking, and a comprehensive keyword index." [8] With these prerequisites met, Luhmann dubbed his creation a "second memory", capable of incidental connections to be made, and new knowledge to be formed by the system itself.

Comparing Luhmann's 'Zettelkasten' with Bush's idea of a 'Memex', we can see many similarities. Although limited due to it's manual nature, the Zettelkasten is an early form of the sort of memory expansion Bush was speaking of in his article. Although not able to store "all his books, records, and communications" [1], it is by definition an "enlarged intimate supplement to [Luhmann's] memory."[1] Moving forward through the advent of new technologies, including the world wide web, It may be possible now to use Luhmann's methodologies in a renewed context.

II. Challenges

In 2005 a group of researchers at the University of Colorado attempted to summarize and evaluate attempts at the Memex through a technical report. To accomplish this task, they not only defined a number of terms that are useful in context to describe the goals of a Memex, but they also surveyed a number of attempts to realize Bush's idea. Through nine design goals, and four overarching design choices, Stephen Davies, Javier Velez-Morales and Roger King examine many attempted solutions of their time, and eventually settle on the fact that there is not yet a solution capable of accomplishing Bush's original intent as adapted to fit our current technological capabilities.

Before continuing, I must note that Davies et al. use specific terminology throughout their paper that I will reference throughout the following section. I will not go into great detail covering all of their definitions, but the following are some of the most important terms summarized:

Personal: A user's subjective realm, whose trends, relationships, categories, and personal observations are truths for an individual, but it may be possible that no one else agrees with.

Knowledge: The distilled version of the particular truth an individual is seeking, such that a mental model once perceived can be easily reformed. The question of "What is knowledge?" can be answered with another question "What have I learned?"

Base: A single unified whole of which knowledge is consolidated, integrated, and connected without explicit partition. The important part of having a 'base', is that it is only ever singular.

Data: Any potential information whose value is not yet determined. Data is potentially information, and only through a standard language, or convention can the information be perceived.

Information: Any material that has the potential to be used for or to become knowledge. Information in its raw form can be thought of as an encyclopedia waiting to be read and whose contents are waiting to be parsed through and understood.

In an attempt to realize Bush's ideations, I will follow Davies et al's recommendations as written in their report, and then explain some differences and changes to design that may be more beneficial to the final result, as they dub "Memex II". [3] To do this, I will first briefly cover their goals and design ideologies, and then explain possible choices and developments to achieve these goals.

III. Design goals & choices

The first tenet Davies et al. proposes is that a Memex must "operate as the brain works" [3] by not only associating

any two topics together without restraints, but doing so semantically rather than syntactically, all whilst keeping the knowledge therein easily accessible and viewable in many ways. Secondly, the 'Memex II' to be useful in the long term must contain knowledge from all domains of a person's life. Segregating a 'work life' from a 'personal life' is in this way an antithesis of the overarching idea of a second brain. Just as humans do not entirely separate ideas between domains, neither should this thinking tool. Furthermore, This tool must both support formality and informality within. What this means is that ideas in any state, whether be pre-processed ideas (raw information not yet processed into core knowledge), key understandings, or external files must all be able to be contained within the Memex. The group of researchers move forward with their goals by explaining that a user of such a tool must be able to easily mutate and refactor any piece of knowledge. This must be accessible enough through both a "streamlined user interface" and "ubiquitous availability", such that it can inspire "a willingness to experiment" with different connections of knowledge and ideas that can be formed. [3] Additionally, and this point may prove redundant as technology has progressed since time of writing, "Immediate retrieval" [3] of previously stored information is also a key tenet of a Memex system. Finally, A memory expansion tool must be able to bridge the gap between objective and subjective information. This bridge means simply to allow knowledge obtained to refer back to the information from which it was derived. In this way, a linking archivable reference to how that knowledge was ascertained may be created.

Following the goals written, Davies et al. continue with certain characteristics that they felt at time of writing that a Personal Knowledge Base should possess. This includes a full transclusatory method for modeling data in a number of ways, including a way to show relationships within a semantic network of knowledge "snippets", images, and other elements. The researchers also double down on the importance of a streamlined user interface, ensuring that it is as simple and free flowing to enter knowledge as it is for a human to think. This simplicity is aimed at encouraging the user to enter as much knowledge as possible, since the opportunity cost and effort required will be very low. The back-end side of a Memex is theorized to be best fit as a being a database, keeping any connections stored alongside the knowledge itself "interconnected in an unconstrained manner."

IV. Creating a Personal Knowledge Base

I have made a few exploratory attempts at realizing the vision as written by Davies et al. My first attempt was closely in line with their vision, a database hosting all of the knowledge, relationships, and files associated. The relational database structure was as seen in Fig 1, with each piece of knowledge being neatly wrapped in it's own 'Tile'. These tiles or 'Unit' each had a user, multiple tags, and had the possibility of relating to a file to source the information behind the knowledge. (See Fig 2.) This was coupled with a web accessible front end, to allow for "ubiquitous availability", and "streamlined user interface", both of which are requirements for a personal knowledge base. (See Fig 3.) [3] Once the knowledge was stored in the database through the web portal, it would then later be accessed in a synonymous fashion to Luhmann's Zettelkasten, wherein the 'topics' you would add to any one tile would represent tags in an associative system, eventually resulting in a system of connections looking like the following image. (See Fig 4.) This was the beginning of a two dimensional semantic map of which one would be able to view and interact with their knowledge. Functionally the associative map is the visual realization of the "web like network" Luhmann had created with his "Zettelkasten" system.

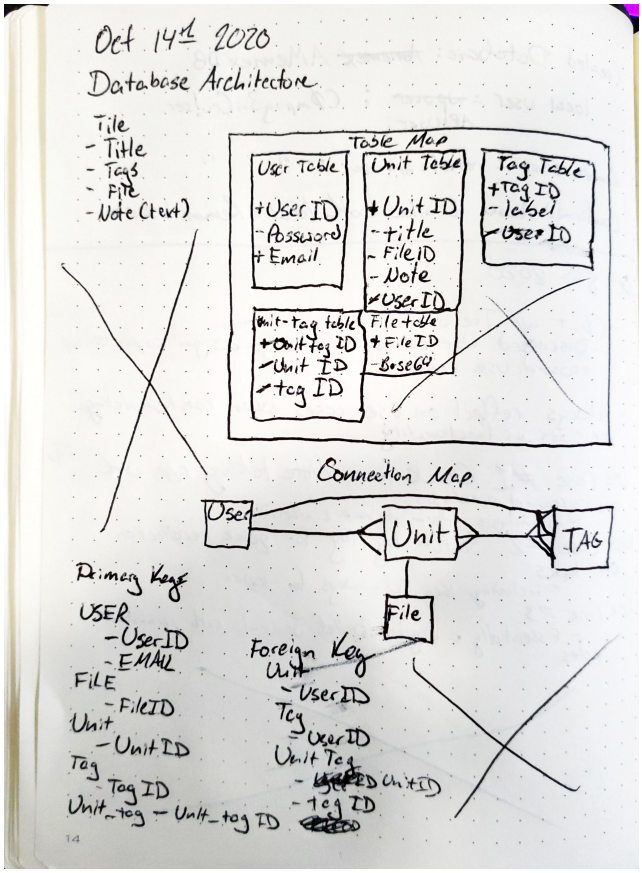

Fig. 1. Potential Database Architecture

A relational database isn't without its flaws. Although the infrastructure allows for accessibility across the network, the way in which the data is stored is restrictive, and does not easily allow for "heterogeneous sources"[3], nor is it in general flexible enough invoke curiosity, or the willingness to experiment with any one piece of knowledge, as the 'tiles' in this format make the knowledge feel more segmented than they are in reality. Instead, I moved forward exploring a possibility that conversely to Davies et al's tenets, a file based approach may in fact be the best approach.

At the time of writing this report there is one application I have found that is able to suffice nearly all of the tenets

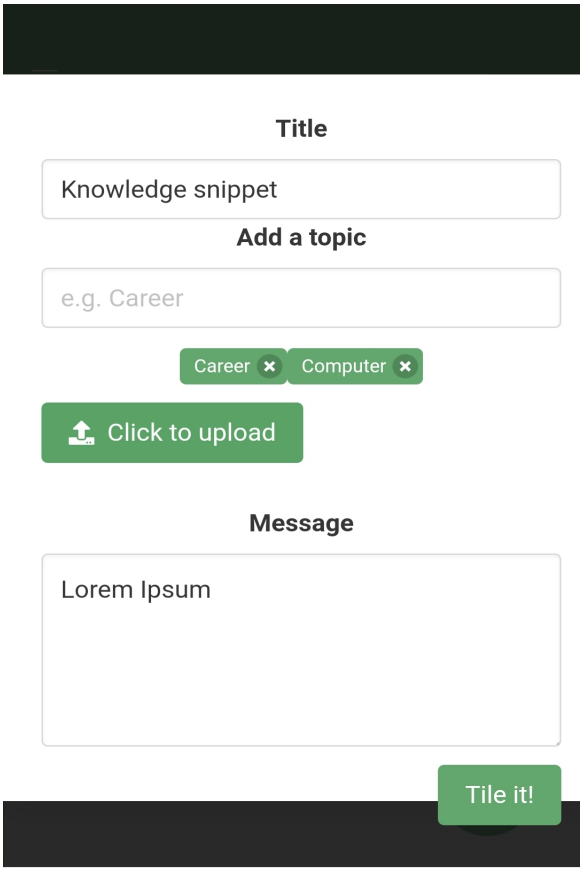

Fig. 2. Tile Input View

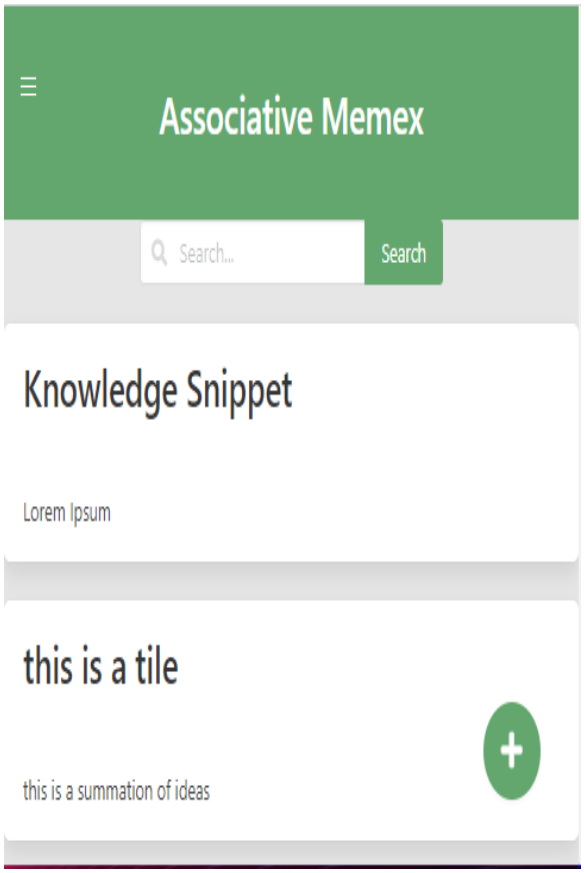

Fig. 3. Potential Default View

Davies et al. put forward in their 2005 report. Obsidian.md is "a powerful knowledge base that works on top of a local folder of plain text markdown files." [5] This approach allows obsidian to use certain commonplace syntaxes within a plaintext file to facilitate associative connections required for a note taking app to become "a second brain" [5].

This is a more dynamic approach than storing associations in a database, allowing transactions with data to happen as quickly and as readily as you can open a file. Furthermore, this allows for a few key differentiators that any other approach will lack. This first feature that a file-based approach flexes, is the ability for any front end document editor to be used when entering knowledge. A second feature is that the simple file structure allows all of the knowledge to be quickly replicated, duplicated, and hosted anywhere. Obsidian.md at time of writing this report is still in a public access beta test, however as it stands I believe it can be compared to the tenet's of a Memex and may be thought of as a successful attempt, with a few caveats.

As it currently exists, Obsidian works with local files. This inherently poses the problem not only of accessing the files anywhere and anytime being an impossibility, but also removes any idea of immediate retrieval being possible unless you are physically able to manipulate the device that these files exist upon. In 1945 Bush proposed the idea that a Memex be a stationary device that one must return to

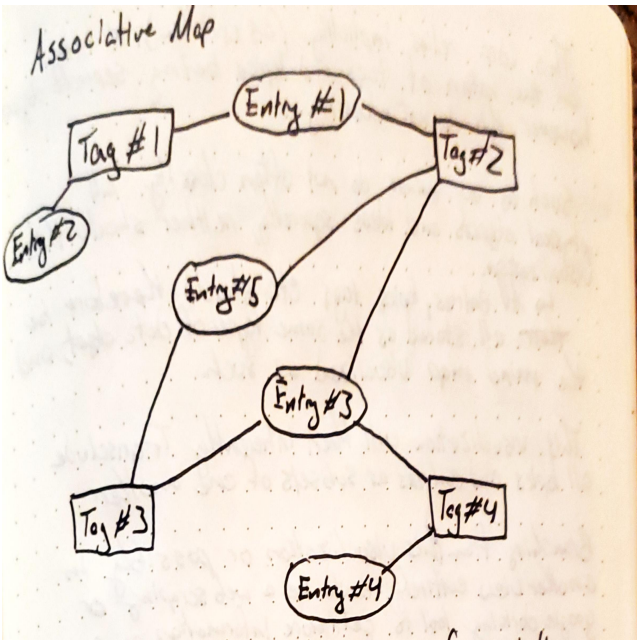

Fig. 4. Webbing of Ideas and Knowledge via Tags

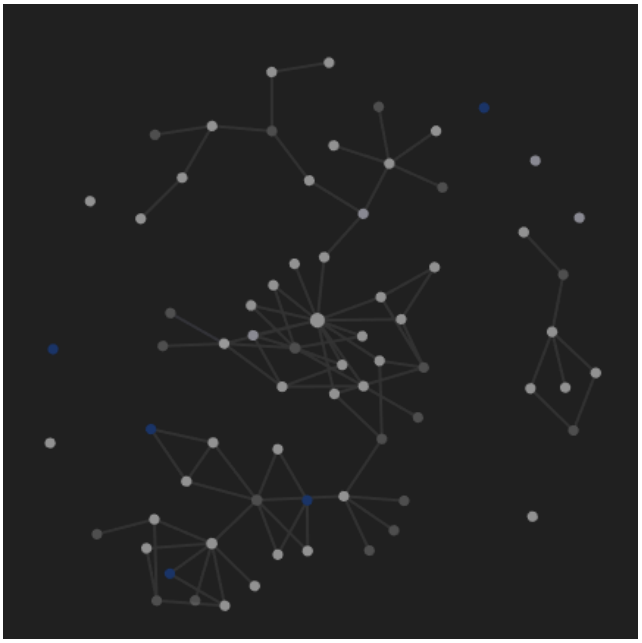

Fig. 5. Webbing Knowledge Points in Obsidian

before accessing. As Davies et al. eluded to in their 2005 article, technology has surpassed the idea of static files, and instead accessing a personal knowledge base must be possible from a "mobile device" [3]. Solving the problem of accessibility is in fact a more simple problem than I had initially expected. Given that obsidian recursively searches through every subdirectory in a folder, one only needs to make a folder web accessible to enable "mobile device" access. I have identified two ways to go about this: one is to host the folder on a web accessible file server, and the other is to make use of a third party repository host such as "Github" (https://github.com/), and constantly synchronize any client of which you wish to interact with the Memex on.

The second caveat I have identified is a problem that every attempt at a "second brain" [5] software has encountered to date. These softwares do not attempt to discover associations ad hoc, and instead rely on associations the human has made explicitly or implicitly. I propose that through the use of natural language processing (NLP), it is possible to have a computer program look through a file based system, and analyze the text that exists for possible tags and topics that the human did not otherwise denote. Through the use of NLTK, a natural language toolkit, and Gensim, a library focused on topic modelling, I have developed a python script capable of scrubbing through markdown files, and inserting topics with syntax such that obsidian is able to model associations between not only knowledge that a user has inputted, but also associations between raw information. The associations that can be automatically created between otherwise unparsed information creates dynamic opportunity for human intervention, and knowledge creation. (See Appendix I) This association creation is what I believe to be a key factor in a complete realization of Bush's Memex.

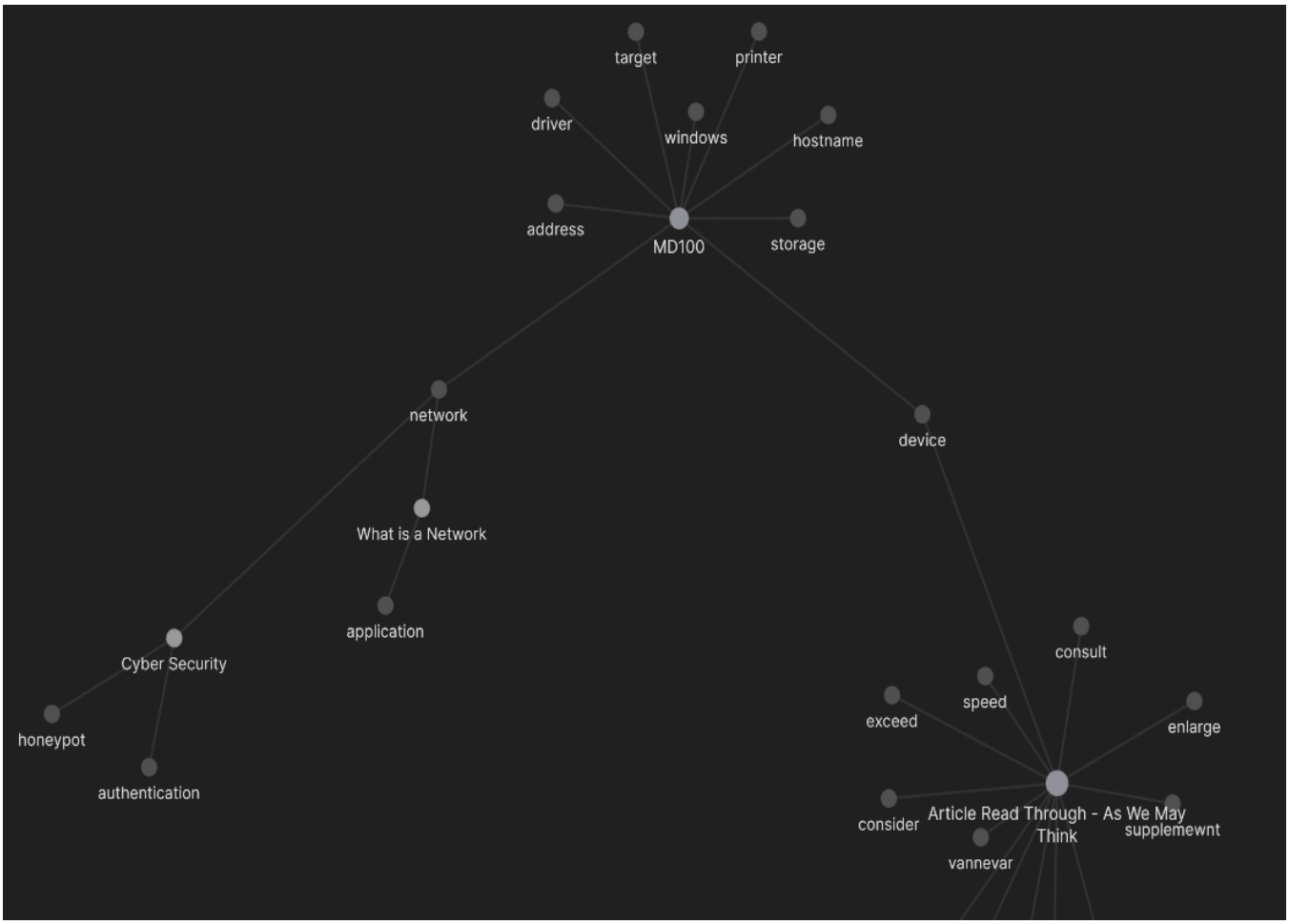

Fig. 6. Associative Connection in Obsidian

Seen in Fig. 6 is a dynamic association created between two sources of ‘information’ and one piece of ‘knowledge’. These pieces of information are two sources that I have copied into the file structure as markdown files, whereas the source of knowledge is my summation of thoughts regarding a course I have taken in the past. In this example, both ‘What is a Network’ and ‘Cyber Security’ are informational articles written on topics as their name would suggest, and they have been found to relate to the dissemination of knowledge I have created regarding the Modern Desktop course I have taken in my past. The semantic relationship found has the potential to be noted explicitly by the human involved, but in this case was found automatically by the natural language processing script. After allowing for a few ‘passes’ of the articles, the script found the following ‘tags’ for the information, and for the knowledge.

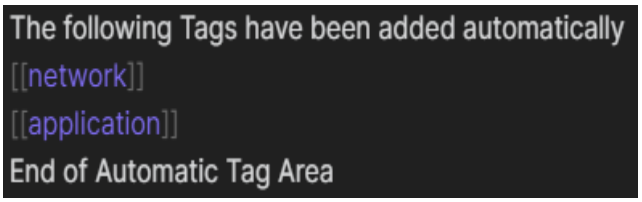

Fig. 7. ”What is a Network” tags

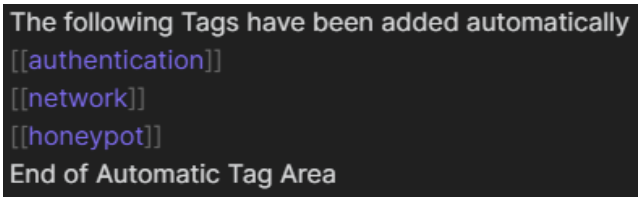

Fig. 8. ”Cyber Security” tags

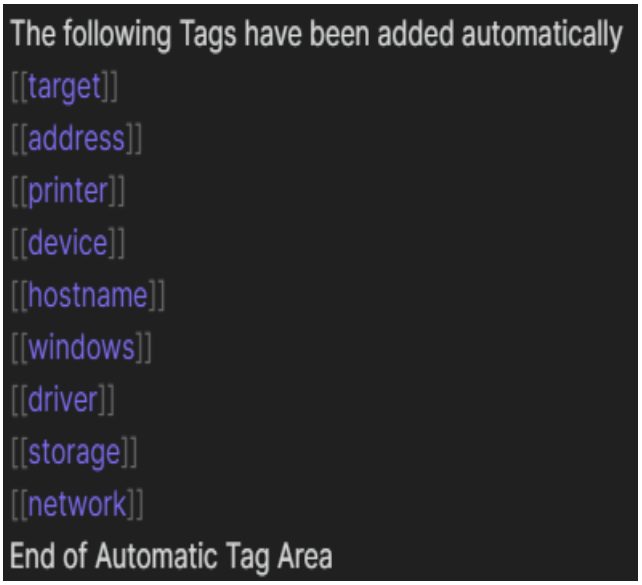

Fig. 9. ”MD100” tags

As seen in Fig 7, Fig 8, and Fig 9, ‘network’ is an explicit connection between each of these items. As such it can be expected that the user has some knowledge about networks. I propose that this such example is one bridge between the “objective and subjective” domains that Davies et al. sought to obtain.

Many connections created this way will be rather obvious to the human involved, and such is the case with ideas normally created. There will however be stretch cases where new knowledge is identified that is not immediately obvious to the human that wrote their pieces of knowledge. In this way, connections can be spawned as a ‘new beginning’ for knowledge to be created; Oftentimes, this new knowledge is highly personal and may only make sense to the owner of the Memex. In this example below, the NLP performed on my personal knowledge base has made a connection between the lore of a fictional dungeons and dragons character, and my collection of notes regarding negotiation tactics. In Fig. 10 we see the semantic connection has been made through the use of the English term “party”.

V. SUGGESTED IMPROVEMENTS

The Memex I have proposed is a combination of Obsidian.md, and some python scripts. It is far from a perfect solution. Some additions that could be beneficial would be auto source gathering. This can be done through querying the world wide web for scholarly articles, or simple querying the internet for key terms. Automating the web searching process and including a way to link source documents via the same internal linking style that obsidian already supports could be beneficial. Furthermore, support for image uploads with image recognition software automatically associating it with tags in the Memex could be incredibly useful for quick uploads whilst you are on the go.

VI. CONCLUSION

Obsidian when coupled with NLP successfully assumes the role of a modern “Memex” as was once coined by Vannevar Bush in 1945. Through the modern age of digitalization the nature of a Memex has changed from being strictly a consultative device, to becoming mobile, dynamic, and “ubiquitously accessible”. [3] Using associative methods popularized by Niklas Luhamnn, it is possible to associate terms though natural language processing. Due to the nature of file based storage, it may be possible in the future to include automatic source capturing, as well as other features in the Memex. Bridging the gap between information and knowledge, as well as the objective and subjective domains, the Memex can be used by researchers, students, and anyone else looking to keep a permanent and dynamic record of their personal knowledge.

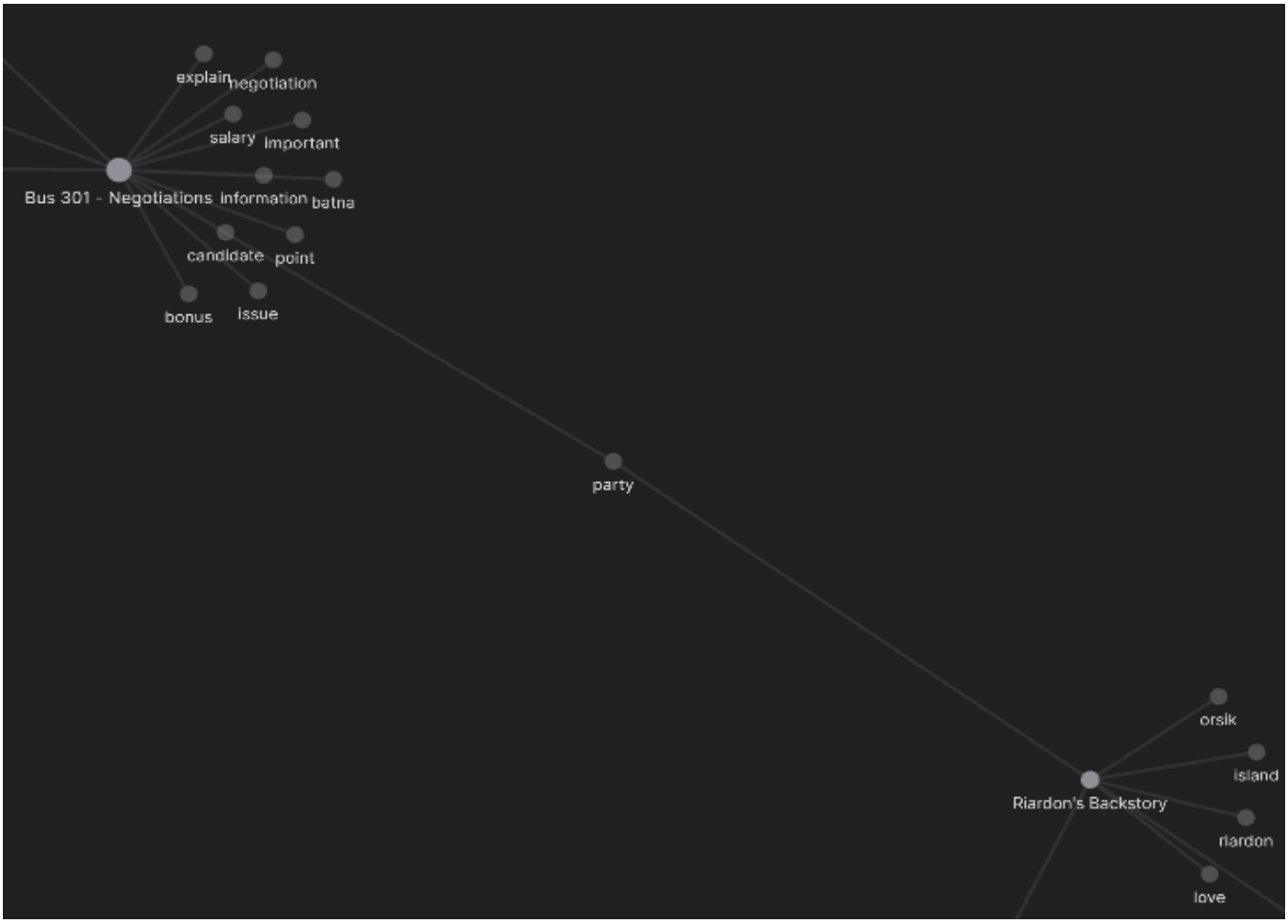

Fig. 10. Semantic associations created automatically

REFERENCES

[1] Bush, V. (1945, July). As We May Think. The Atlantic, (July 1945), 101-108.

[2] Kapadia, S. (2019, September 05). Topic Modeling in Python: Latent Dirichlet Allocation (LDA). Retrieved November 13, 2020, from https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

[3] King, R., Davies, S., &; Velez-Morales, J. (2005). Building the Memex Sixty Years Later: Trends and Directions in Personal Knowledge Bases ; CU-CS-997-05. Retrieved 2020, from https://scholar.colorado.edu/concern/reports/t722h9830

[4] Li, S. (2018, April 03). Topic Modelling in Python with NLTK and Gensim. Retrieved November 13, 2020, from https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21

[5] Li, S., Xu, E. (2020). A second brain, for you, forever. Retrieved November 13, 2020, from https://obsidian.md/

[6] Linvega, L. (2020). XXIIVV - indental. Retrieved September 14, 2020, from https://wiki.xxiivv.com/site/indental.html

[7] Luhmann, N. (1951-1996). Zettelkasten Niklas Luhmann. Retrieved 2020, from http://ds.ub.uni-bielefeld.de/viewer/toc/ZKLuhm/

[8] Schmidt, J. (2018). Niklas Luhmann's Card Index: The Fabrication of Serendipity. Retrieved 2020, from https://doi.org/10.6092/issn.1971-8853/8350