UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI
DEL TERRITORIO E DELLE COSTRUZIONI

RELAZIONE PER IL CONSEGUIMENTO DELLA
LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE

Mining Healthcare Patents: New Methods and Applications to Diabetes Devices

SINTESI

RELATORI

Prof. Ing. Gualtiero Fantoni, PhD
Dipartimento di Ingegneria Civile e Industriale

Dr. Filippo Chiarello, PhD
Dipartimento di Ingegneria dell'Energia,
dei sistemi, del territorio e delle costruzioni

Dr. Vito Giordano
Dipartimento di Ingegneria dell'Informazione

IL CANDIDATO

Greta Giannecchini
giannecchini.greta@virgilio.it

Sessione di Laurea Magistrale del 17/02/2021

Mining Healthcare Patents: New Methods and Applications to Diabetes Devices

Greta Giannecchini

Sommario

I brevetti contengono importanti informazioni che possono rivelarsi essenziali per diversi *stakeholders*, quali policy maker, università e aziende. Se analizzati attentamente, possono mostrare dettagli e relazioni tecnologiche, rivelare tendenze di business, ispirare nuove soluzioni industriali o aiutare a fare politiche di investimento.

Il presente lavoro di tesi ha lo scopo di utilizzare questi documenti come strumento per mappare i concetti relativi all'ambito medico. In particolare quando si tratta di malattie croniche come il diabete è di fondamentale importanza lo sviluppo di tecnologie in grado di migliorarne la gestione. Molti progressi infatti, sono stati recentemente raggiunti, inglobando concetti come *Industria 4.0* e *IoT* all'interno delle invenzioni mediche. Nella mia tesi, attraverso lo studio dei brevetti è stato possibile analizzare gli sviluppi tecnologici per gestire il diabete, indirizzati sempre di più verso l'utilizzo di tecnologie *smart*. Sono stati utilizzati strumenti di analisi automatica del testo per estrarre tre tipologie di entità: gli utenti citati nelle invenzioni, le malattie correlate al diabete e le tecnologie riferite all'Industria 4.0. L'analisi di queste tre entità ha permesso di capire su quali dispositivi si sta orientando la ricerca, quali sono le malattie su cui ci si concentra maggiormente e a quali tipi di utenti saranno indirizzati.

Abstract

Patents contain important information that can be essential for different *stakeholders*, such as policy makers, university and companies. If analyzed carefully, they can show technological details and relationships, reveal trends, inspire new industrial solutions or help in making investment policies.

This thesis aims to use these documents as a tool to map concepts related to the medical field. Especially in the case of chronic diseases such as diabetes, the development of technologies that can improve their management is of particular importance. Through the study of patents, it was possible to analyze the technological developments to manage diabetes, which are increasingly directed towards the use of *smart* technologies. Text mining tools were used to extract three types of entities: users mentioned in the inventions, diabetes-related diseases and technologies related to *Industry 4.0*. The analysis of these three entities made it possible to understand which devices are being researched, which diseases are being focused on and which types of users will be targeted.

1 Introduction and scope

Patents are more than just legal documents, they contain a large amount of information that can be essential for different stakeholders such as policy makers, researchers and private companies. If analyzed carefully, they can show technological details and relationships, reveal business trends, inspire new industrial solutions or help make investment policies. Several studies[1] [2] have shown that about 80% of the technical information contained in patents is not available elsewhere, so patents are one of the most comprehensive resources for technical analysis. Chiarello et. al[3] show that, in addition to technical information, a fraction of all other types of information are not contained anywhere else, as early publication of information about the invention may compromise its patentability. Therefore, the constant monitoring and consultation of patents should be of interest not only to patent attorneys, but to all those who need to know the state of the art in a particular field and to keep up with the latest technological developments.

Thanks also to the development of Text Mining, analyzing patents no longer involves great human effort, so identifying information within a large body of text has become easy and time-saving.

Therefore, this thesis work aims to use patents as a tool to analyze and investigate the medical domain. In particular, patents related to diabetic devices have been taken as a reference. Starting from the state of the art on diabetes, it is highlighted how the devices used for its management and treatment are increasingly focused on connection and interaction with other technologies, data exchange and telemedicine. So, the aim is to automatically map concepts related to diabetes care, and especially to understand how diabetes-related inventions have evolved over time, reconciling increasingly smart and mobile technologies. To do this, NER (Nominal Entity Recognition) techniques were used to automatically extract three different types of entities: technologies related to Industry 4.0, to understand if indeed inventions are moving towards participatory and personalized healthcare, ensuring data is transmitted to different actors by interconnecting different devices; the users mentioned in the inventions, to analyze who are the end users of the

[1] *Quick Scan: a novelty search service in the framework of Euro-R&D programmes.* W.Kütt, M.Schmiemann. 1998, World Patent Information, Vol. PC-22, pp. 146-147

[2] *Patent as Technical Literature.* P.J.Terragno. 2, s.l. : IEEE Transaction on Professional Communication, 1979, Vols. PC-22

[3] *Automatic users extraction from patents.* Filippo Chiarello, Andrea Cimino, Gualtiero Fantoni, Felice Dell'Orletta. s.l. : World Patent Information, 2018.

devices and with whom they interact; diseases, to understand how diabetes inventions have evolved: if before devices were created specifically for diabetes, now the scope has expanded, incorporating also co-morbidities and complications or even diseases that can be treated using the same technologies.

2 State of Art

In order to set the context, an initial phase of this thesis was a thorough review of the scientific literature relating to diabetic topics and the use of Text Mining in the medical field.

2.1 Diabetes Mellitus

Diabetes Mellitus (DM) is a chronic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. DM is one of the most common endocrine disorders, affecting more than 400 million people worldwide. Due to high DM mortality and morbidity as well as related disorders, prevention and treatment attracts broad and significant interest. In recent years, research has focused on the use of mobile health technologies to help patients better manage diabetes and improve their quality of life.

2.2 Text Mining in Healthcare

Text mining is widely used in the medical field [4], precisely because the biomedical literature is growing exponentially, producing a large volume of health data that can be used to extract relevant and meaningful information from various sources (scientific publications, clinical trials, physicians' notes, web searches, patents, forums, etc.) in order to improve health care delivery, enhance research, and reduce health care costs.

However, it has been noted [5,6] that Text Mining is mainly used for scientific articles in biomedical literature and clinical information found in clinical information systems, whereas it has not reached the same level of maturity for biomedical patents. In addition, most of the existing work has focused primarily on the detection of genes, proteins, drugs, and anatomical parts.

[4] *A systematic review of text mining approaches applied to various application areas in the biomedical domain*. S Cheerkoot-Jalim, KK Khedo, Journal of Knowledge Management, 2020, 1367-3270

[5] *A brief survey of text mining.* Hotho, A., Nürnberger, A. et al, in Ldv Forum, 2005, Vol. 20, pp. 19-62.

[6] *Text Mining patents for biomedical knowledge.* Raul Rodriguez-Esteban, Markus Bundschus. 6, s.l. : Drug Discovery Today, 2016, Vol. 21, pp. 997-1002.

As now widely demonstrated, patent analysis is one of the most effective methods to understand technology trends[7], to measure their maturity[8], as well as to identify new technological and commercial opportunities[9]. Moreover, patents contain not only technical information about the invention they deal with, but also information about stakeholders and information about needs it manages to satisfy.

3 Methodology

The process used in the methodology is divided into three macro parts:

1. *Patent Retrieval;*
2. *Entities Extraction;*
3. *Knowledge Discovery*

3.1 Patent Retrieval

The aim of the first methodological step is to retrieve all and only patents that are related with diabetic devices. A query was created to retrieval the patent documents on Errequadro s.r.l. database. This database is a proprietary database containing over 90 million patents from the DOCDB and European Patent Office (EPO) repositories. The ability of the query to retrieval only relevant document is manually validated and the query precision is calculated.

3.2 Entities Extraction

After the relevant documents have been retrieved, search algorithms are used to analyze the documents for the identification of specific keywords and relationships between them. The aim of this phase is to identify for each document the technologies 4.0, users and diseases mentioned in each patent of the diabetic devices. A Text Mining technique, called as Nominal Entity Recognition (NER), enables to automatically collect the three types of entity. The NER approach used in this thesis work (Gazetteer NER) aims to map mentions of entities within texts using a list of entries. The input list of users was collected by taking the list of users produced by Errequadro s.r.l[10], the input list of technologies 4.0 was collected by

[7] *Patent analysis for analysing technological convergence.* Karvonen, M., Kässi, T., 2011. Foresight 13, 34–50

[8] *Identification of the technology life cycle of telematics: a patent-based analytical perspective*. Chang, S.-H., Fan, C.-Y., 2016. Technol. Forecast. Soc. Chang. 105, 1–10.

[9] *A novel approach to forecast promising technology through patent analysis*. Kim, G., Bae, J., 2017. Technol. Forecast. Soc. Chang. 117, 228–237.

[10] *Automatic users extraction from patents.* Filippo Chiarello, Andrea Cimino, Gualtiero Fantoni, Felice Dell'Orletta. s.l. : World Patent Information, 2018.

considering the list of regexes within the *Tecnimetro 4.0*[11], while the list of diseases was collected by taking the *ICD-11*[12] (International Classification of Disease, 11th revision) site as a reference. Once the entities had been extracted, a manual cleaning was carried out.

3.3 Knowledge Discovery

The trends of the three entities over time were analyzed, considering the number of times the entities were mentioned in patents per year, in order to understand, for each of the three entities, how many different types are cited.

Then, the following activities were carried out: i) *Normalization*: users and technologies that were spelled similarly (e.g. plurals and singulars of the same word) were grouped together. To do this, algorithms measuring the string distance were used[13]; ii) *Classification* of users and diseases: for users a manual individuation of the main categories was carried out, while the extracted diseases were brought back to the macro-class defined by ICD-11; iii) *Analysis* of the different entities, counting mentions within patents.

4 Results

4.1 Patent set creation and validation

The query adopted in the Errequadro's database generated 4970 results, which filtered by DOCDB families became 2382. A total of 1621 patents were manually analyzed for validating the query, as described in Section 3.1. The number of relevant patents is 1405 on 1621, thus the precision[14] of the query is 87%.

Finally, an analysis of the patent set was carried out to better understand the knowledge produced in the field of diabetes devices. Figure 4.1 shows the number of diabetic device patents produced per year and the cumulative curve[15]. The fact that the curve assumes this upward trend is influenced by the general growth in scientific discoveries and especially by the growth in patent activity. However, it can be seen that the number of patent families, with an average of 56 patents per year for the years 2002-2007, increased more than

[11] *Extracting and mapping industry 4.0 technologies using Wikipedia.* F. Chiarello, L. Trivelli, A. Bonaccorsi, G. Fantoni. s.l. : Computers in Industry, 2018, Vol. 100, pp. 244-257.

[12] https://icd.who.int/browse11/l-m/en

[13] https://cran.r-project.org/web/packages/stringdist/stringdist.pdf

[14] The precision is the fraction of retrieved documents that are relevant to the query. See https://en.wikipedia.org/wiki/Precision_and_recall

[15] The main y-axis (left) refers to the cumulative curve, while the y-axis on the right is the secondary one, relating to the number of patents per year.

threefold for the years 2014-2018. Indeed, as A. Vesselkov et al.[16] states, *"the share of patents on glucose measurement has grown particularly rapidly, making diabetes management one of the most developed applications of telehealth at the Invention level."*

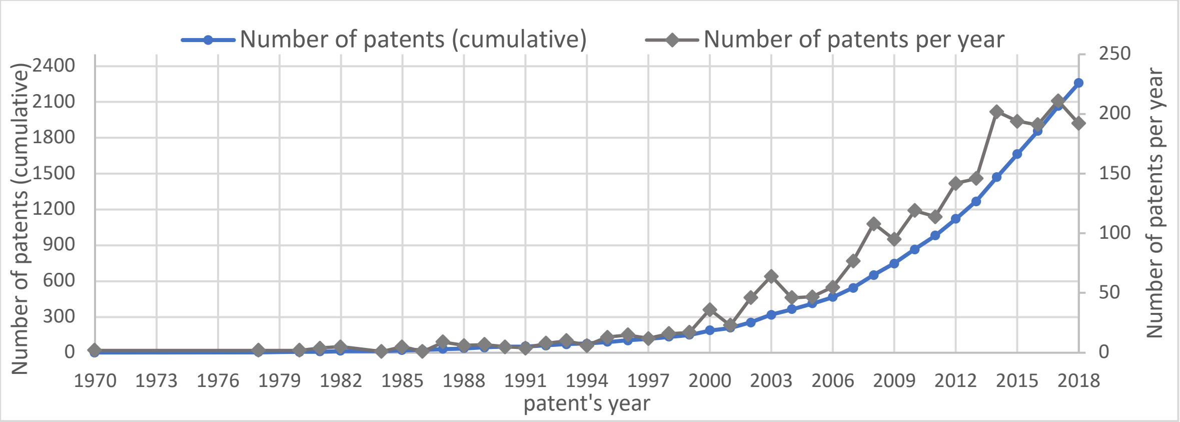

Figure 4.1 Number of patents on diabetic devices produced per year (grey) and the cumulative curve (blue)

4.2 Extraction and Analysis of the Entities

This chapter shows the results obtained by analyzing the extracted entities. In Paragraph 4.2.1, 4.2.2 and 4.2.3 the results of the various entities analyzed separately will be shown.

Figure 4.2 shows the trend of the three extracted entities over time: each curve represents, for each year in which a patent was published, the total number of different entities present in the patents of that year (not considering duplicates per patent). Through this graph, it can be seen that the entities have increased in terms of types in recent years, meaning that more and more different entities are contained within a patent. The fact that the curves show an increasing trend is certainly conditioned by the trend of the patent curve: the three curves in fact increase for the same points as the patent curve (Figure 4.1).

Another important observation concerns the gap between the three curves (technologies, users and diseases). In fact, it can be seen that the entities most present in patents are technologies 4.0, followed by users (it is important that within patents it is specified to whom the invention is useful) and finally diseases (representable as users' needs).

[16] *Technology and value network evolution in telehealth.* A. Vesselkov et. al., 134, s.l. : Technological Forecasting & Social Change, 2018, pp. 207–222.

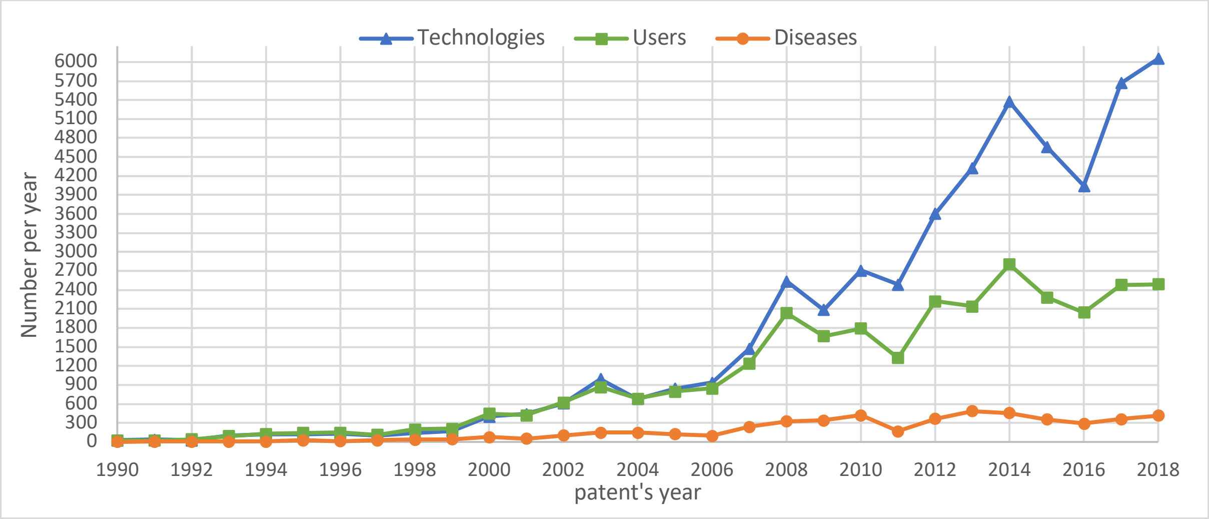

Figure 4.2 Technologies 4.0 curve (blue), users curve (green) and diseases curve (orange)

4.2.1 Extraction and Analysis of Users

The number of users extracted was 32.366, resulting in 745 different types.

By analyzing the extracted, cleaned and normalized users, the most commonly occurring classes in the patent set were identified. The first 6 classes that appear most frequently in patents are presented below, in descending order:

Health personnel: this class includes all users referring to health professions (physicians, doctors, cardiologists, diabetologists, nurses, etc.). These users are mentioned in patents because they interface with the device used by the patient (e.g., the device collects data that is then sent to and analyzed by physicians).

Patients: all extracted users referring to sick people (e.g. "diabetes mellitus patient" or "paralyzed" or "suffering from heart disease"). Generally, these users represent the end user of the device that treats the disease or improves its management.

Animal species: all extracted users referring to animal species (e.g. "dog", "chimpanzee", "monkeys"). These users are found in patents on medical devices both because very often animals are used to test new inventions and because many inventions can also be used to treat animals.

Relatives: encloses users such as “brother”, “grandparents”, “mum”, “wife”, “father”, “brother”. Users of this class appear in the patents as support for the user of the device. They usually interface with the device themselves, e.g. if the patient is a child or elderly (and therefore unable to use the device themselves).

Patent field: users referring to the patent field ("applicant", "skilled artisan", "skilled in the relevant art”). Users in this class are considered outliers in this analysis, but are mentioned because these users are always present in the text of the patents.

Age-related: extracted users who had an age reference, such as “adolescent”, “children”, “adult”, “old people”, “teenager”, “toddler”.

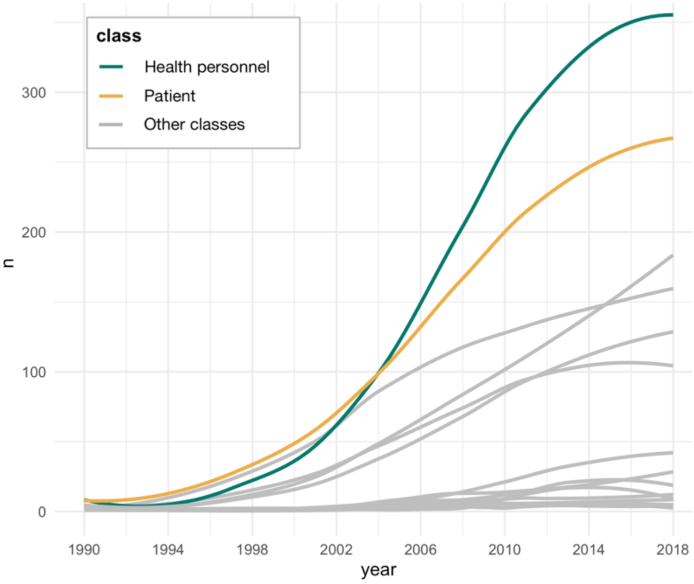

Figure 4.3 Development of users classes over time

The graph in Figure 4.3, where the x-axis represents the year of publication of the patents and the y-axis represents the number of patents containing the classes, shows the changes that the classes Patients and Health personnel have undergone. In particular, patents used to focus more on patients, whereas since 2004 the trend has reversed, focusing more on health professionals. This means that the scope of application of devices for diabetics has been widened, involving a greater number of other health professionals (caregivers, medical specialists, nurses, etc.).

This gives evidence of the development of telemedicine in diabetic inventions. In fact, telemedicine allows an interactive relationship between doctor and patient, but not only: the interaction also takes place between general practitioner and specialist doctor, between these and the nursing staff, both inside and outside the health facilities, thus allowing the improvement of the health service and disease management thanks to a greater collaboration between the various health professionals involved and the patients. In fact, a study[17] shows that diabetes was the most targeted application of telemedicine products launched in 2002-2011.

4.2.2 Extraction and Analysis of Technologies

The number of technologies 4.0 extracted was 55.047, resulting in 1.899 different types.

Figure 4.4 shows, for the first 19 technologies 4.0, the total number of patents in which they are mentioned.

“Sensor”, which is the most prevalent technology in patents, is almost always mentioned in connection with glucose monitoring. While the other technologies most often mentioned (computers, software, wireless, etc.) serve to enable the interaction and exchange of data.

[17] *Technology and value network evolution in telehealth.* A. Vesselkov et. al., 134, s.l. : Technological Forecasting & Social Change, 2018, pp. 207–222.

From the results, it is clear that a digital transformation of healthcare is underway, enabling productivity improvements, cost containment and better user experiences.

Devices to manage diabetes (and health in general) are increasingly oriented towards and integrated with Telemedicine and IoT (internet of things).

Figure 4.5 shows the development of some key technologies for the integration of medicine and IoT, such as the use of Bluetooth to connect, for example, the glucometer with a monitoring device, or the use of the Internet to transfer information on diabetes management, with the creation of dedicated portals as well.

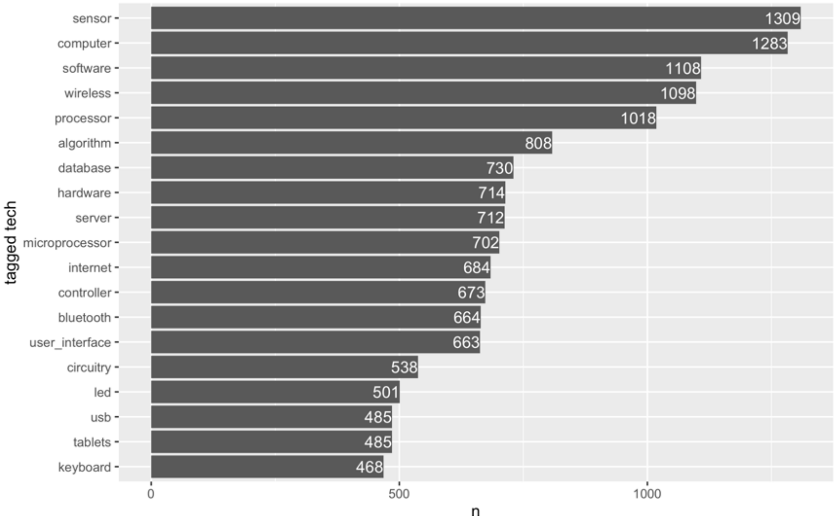

Figure 4.4 Top 19 technologies ordered in terms of occurrence

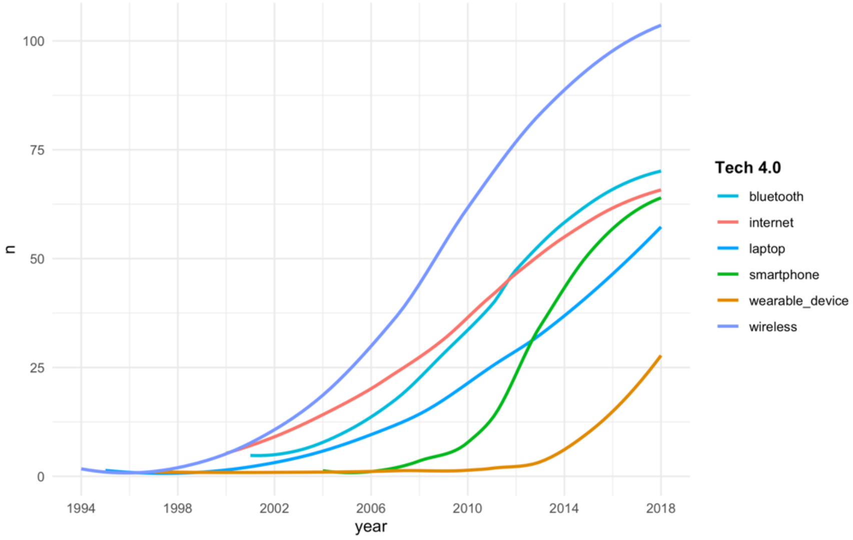

Figure 4.5 Technologies 4.0 trends

These technologies have increased in recent years in terms of mention in patents, highlighting how diabetes inventions are directed towards this phenomenon.

4.2.3 Extraction and Analysis of Diseases

The extraction yielded 5.610 extracted diseases (unique per patent), resulting in 528 different types.

Considering Figure 4.6, the disease curve (orange) has grown. Although this is influenced, as mentioned above, by the increase in patents, the gap between the disease curve and the patent curve is still large, so it means that in recent years diseases have increased in terms of types: more and more different diseases are contained within a patent. This means that, in addition to the original disease, other diseases are also mentioned in a patent.

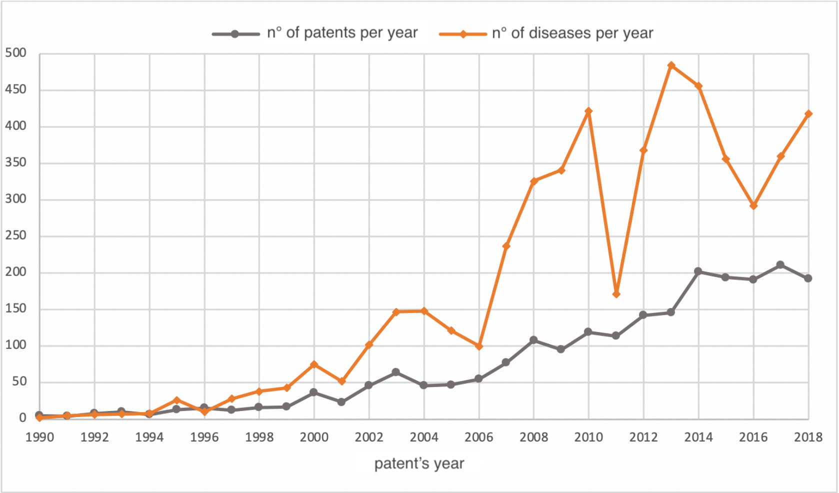

Figure 4.6 n° of diseases extracted per year compared to the n° of patents per year

The increase in the number of diseases and medical conditions covered by patents was also highlighted by Huang et al.[18], suggesting that technological innovations are focusing on more diseases and medical conditions simultaneously. The reason why other diseases appear is twofold: on one hand it means that there is an increase in comorbidity, on the other hand, it may be because a device is trying to manage and treat several diseases, which are not necessarily related, but which use similar devices.

Figure 4.7 shows the 20 most frequent macro-classes in patents, while Figure 4.8 shows the top 12 most frequent diseases in patents, with reference to the class to which they belong. The number shown in the bar expresses respectively the number of patents containing that class and that disease.

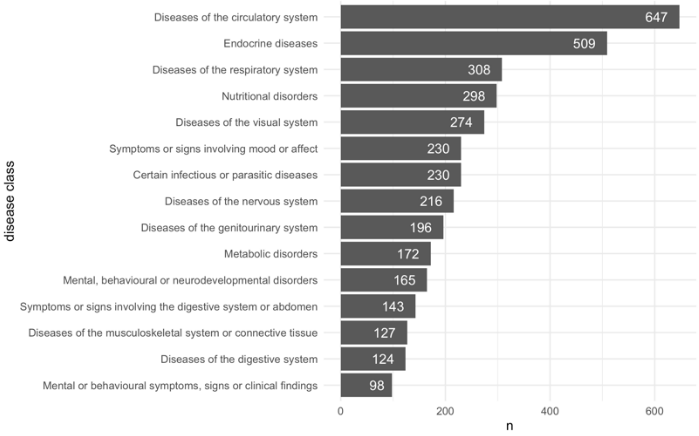

Figure 4.7 Top 20 Macroclasses mentioned

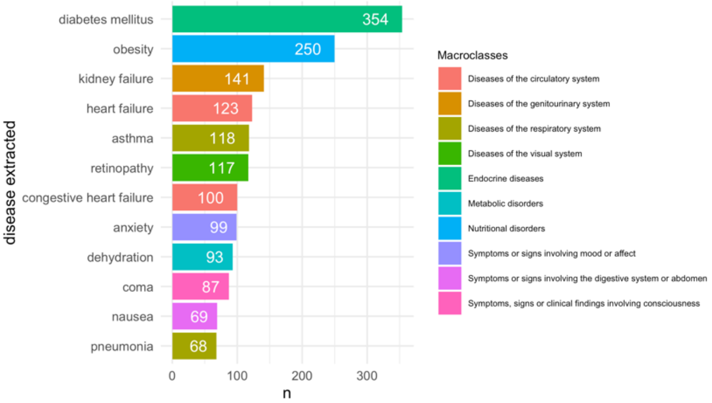

Figure 4.8 Top 12 Diseases with the corresponding Macroclass

As expected, it can be seen from the graph that the most frequent disease in patents is diabetes mellitus. With regard to the other most commonly occurring diseases (with their relative classes), it can be said that most of them are closely related to diabetes: some of them are common complications of diabetes, others are less common, and still others may increase the predisposition to diabetes.

[18] *Technological Innovations in Disease Management: Text Mining US Patent Data from 1995 to 2017.* Huang M., Zolnoori M., et al. 4, s.l. : Journal of Medical Internet Research, 2019, Vol. 21.

5 Conclusions and follow-up works

From the results of the analysis, it can be stated that: i) devices for diabetics increasingly mention different types of users: the focus is on improving disease self-management, so diabetes patents are increasingly addressing the interaction between different types of users, such as caretakers, family, software providers, general practitioners, patients and specialists. ii) by extracting diseases from patents, it is possible to understand which are the diseases most related to diabetes, or in any case the diseases on which research is most focused. iii) by analyzing the most cited technologies, it can be seen that the development of telemedicine and the Medical Internet of Things is increasingly integrated in the innovation of diabetes technologies.

In conclusion, this study applied to diabetes can also be carried out on other diseases of interest, and can benefit several stakeholders: first of all, policy makers, by analyzing this information, can decide where to direct research and funding. Researchers, by analyzing patents, would direct their efforts towards research that is more likely to succeed. Patients would have early access to innovations that improve their quality of life.

Finally, as a next step, it would be interesting to work on the relationships between the various entities. As can be seen from Figure 4.7, I have already started to work on the relationship between users and 4.0 technologies, considering the co-occurrence within the same patent. The nodes are colored according to the type of entity (whether users or technologies), while the size of the nodes is proportional to their number of connections.

By cleaning up the results further and studying the relations between specific entities in a more fine-grained way, interesting clusters could be found.

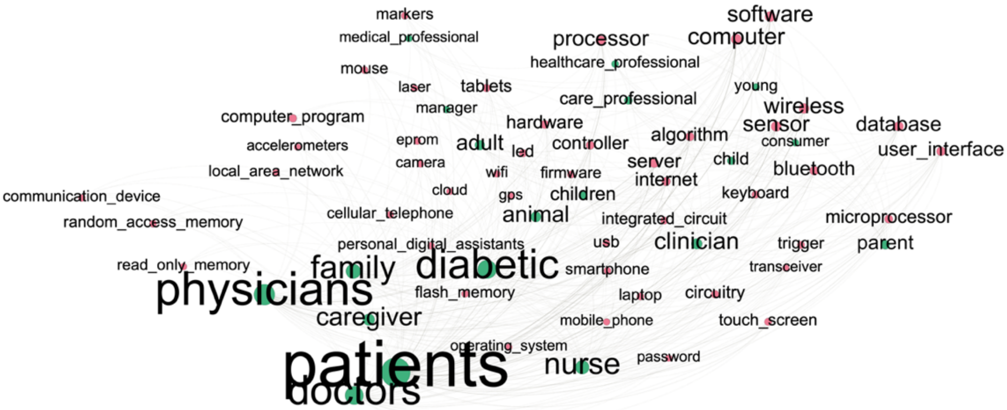

Figure 5.1 Graph representing the connection between users and 4.0 technologies